Improving individual change assessment in clinical, medical and health psychology

Jabrayilov, Ruslan

Publication date:
2016

Document Version
Publisher's PDF, also known as Version of record

Improving Individual Change Assessment in Clinical, Medical and Health Psychology

Ruslan Jabrayilov

Cover Design: Maurik Stomps

Printed by: Ridderprint BV

This research was funded by the Netherlands Organization for Scientific Research (NWO).

Improving Individual Change Assessment in Clinical, Medical and Health Psychology

Proefschrift ter verkrijging van de graad van doctor

aan Tilburg University

op gezag van de rector magnificus,

prof.dr. E. H. L. Aarts,

in het openbaar te verdedigen ten overstaan van een

door het college voor promoties aangewezen commissie

in de aula van de Universiteit

op maandag 4 april 2016 om 14:15 uur

door

Ruslan Jabrayilov

geboren op 11 februari 1984 te Baku, Republiek Azerbeidzjan

Promotor: Prof.dr. K. Sijtsma

Copromotor: Dr. W.H.M. Emons

Overige leden van de Promotiecommissie:

Prof.dr. J.K.L. Denollet

Prof.dr. J.K. Vermunt

Prof.dr. C.A.W. Glas

Prof.dr. R.R. Meijer

Dr. S. Bouwmeester

Table of Contents

Chapter 1

Introduction

There is a growing interest in psychotherapy and counseling among practitioners, researchers, and policy makers. Given limited resources, an important question is whether therapies used in practice are beneficial to patients. This question does not only apply to newly developed therapies, but also to those already used in daily clinical practice. Many new clinical interventions are developed, but their contribution to effective mental health care is not always evident. Furthermore, clinicians have to ascertain that the therapies they use in their daily practice have the desired effects on patients and are therefore advised to continuously monitor their mental well-being. This information is crucial for deciding on the further course of a treatment.

The effectiveness of a therapy is often inferred from within-person change with respect to the intended treatment outcomes across at least two repeated measurements, one measurement before the treatment and one after its completion. A distinction should be made between the mean effectiveness of a treatment at the group and individual levels. At the group level, change assessment entails the comparison of group means before and after treatment. The problem with this approach is that group means may hide important information about the variability of treatment effects on individual patients in these groups. For example, if the mean anxiety level is significantly lower in the treated group than in the untreated group, it is unlikely that each treated individual is feeling significantly less anxious than untreated individuals. Hence, a clinical intervention that works well on average in the population may be ineffective for some individual patients. The reverse may also be true; that is, clinical interventions showing modest average effects may be highly beneficial to some individual patients. Therefore, Jacobson and Truax (1991) argued that the effectiveness of therapies should also be evaluated at the individual level; that is, the percentage of the patients that show a convincing and clinically relevant change should be taken into account. Such individual-level change assessment is also useful to monitor how patients respond to the treatment. Research has shown that regular feedback in the course of a treatment

improves its effectiveness (e.g., Shimokawa, Lambert, & Smart, 2010; see also, Boswell, Kraus, Miller, & Lambert, 2013).

This thesis reports the outcomes of four psychometric studies on various aspects of change assessment in individual patients. For the most part of this dissertation, we adopted Jacobson and Truax's (1991; denoted JT hereafter) methodology to assess change within individuals. The JT method consists of two steps. First, the clinician has to ensure that change between a pretest and a posttest score is real and does not result from random fluctuations caused by measurement errors in a test. This is the test of *statistical significance* of change. The second part of the JT method consists of testing whether a patient's pretest score has moved from the dysfunctional range at pretest into the functional range at posttest, where the functional and dysfunctional populations are defined by clinical cutoff scores. This is JT's test of *clinical significance* of change, which is related to the patient's experience of the meaningfulness of change with respect to the condition from which he or she is suffering. In addition to JT's approach, other approaches to operationalize clinical significance of change have been proposed. One popular approach is to define the minimal change a patient must show before change can be considered clinically meaningful; this is the minimum clinically important difference (MCID) method (e.g., Copay, Subach, Glassman, Polly, & Schuler, 2007; Norman, Sloan, & Wyrwich, 2003). For example, as a rule of thumb many clinicians and researchers consider half a standard deviation to be the MCID for judging change scores to be clinically significant. In this thesis, we also assess clinical significance of change by means of the MCID method.

The JT method was originally defined and used within the framework of classical test theory (CTT; e.g., Lord & Novick, 1968). An alternative framework for test scoring and change assessment is item response theory (IRT; e.g., Embretson & Reise, 2000; Reise & Haviland, 2005; Thomas, 2011). There are several important differences between CTT and IRT, which has led some researchers to criticize using CTT for individual change assessment (e.g., Prieler, 2007). First, in CTT change is assessed based on total scores giving equal weight to all items, whereas in IRT the items are weighted differently when scoring individuals. Second, CTT and IRT treat measurement error differently. In the CTT context, one uses the standard error of measurement to assess the precision of measurement for all individuals, whereas in IRT the conditional error of measurement is used which varies across persons, thus acknowledging that using the same instrument some individuals are measured more precisely than others.

Because JT's test of statistical significance uses the estimated measurement error variance, depending on whether one uses CTT or IRT, inferences about change in individuals can also differ.

Despite the optimism about IRT over CTT (e.g., Prieler, 2007; Reise & Haviland, 2005), the following issues need to be taken into account before deciding which method is better in change assessment. First, there are different methods for estimating person parameters (e.g., maximum likelihood, weighted maximum likelihood, Bayesian methods; see Baker & Kim, 2004, for an overview) which may produce different test-scoring results. These person-parameter estimation methods have been evaluated with respect to the bias in the estimates, but for change assessment it is equally important to know whether the methods also provide accurate and consistent estimates of the standard errors of person parameter estimates. This issue is particularly important when measurements are obtained using a limited number of items (e.g., Magis, 2014).

Second, IRT applications to change assessment require specific psychometric expertise and the use of specialized software. This can be daunting for those lacking the necessary background to apply IRT in practice. In addition, there is a gap between theoretical IRT expositions, however useful, and practical data analysis, although some exceptions are noteworthy (e.g., Brouwer, 2013; Sijtsma, Emons, Bouwmeester, & Nyklíček, Roorda, 2008). This gap may explain why IRT methods are still not mainstream despite their promising features. On the other hand, from a practical point of view, given its simplicity one may also argue in favor of the CTT compared to the more technical IRT. Even though based on theoretical considerations (e.g., Lord, 1980) IRT can be argued to outperform CTT, given that many decisions in psychological practice require only a dichotomous rather than a fine-grained choices along the whole latent attribute scale, measurement precision is not always required to be high throughout the whole scale. This may be one of the advantages of simple CTT methods. Nevertheless, deciding which method to use for test-scoring and change assessment can be overwhelming for practitioners and a considerable part of this thesis is dedicated to studying the differences between IRT and CTT with respect to change assessment.

Another important issue one needs to consider when assessing change based on pretest and posttest scores is whether the measurement instrument has invariant psychometric properties across measurement occasions. Lack of measurement invariance

may involve conceptual changes (gamma change) or a change of the scale metric (beta change). When measurement invariance is violated, making decisions about change based on pretest and posttest scores can be misleading, because the test might be measuring different attributes at different time points, or measurements may be performed on different scales rendering their interpretation problematic. An analogy from physics is a weight scale which provides measurements in kilograms at one time and in pounds at another. One cannot take the difference between the two measurements to assess possible change in weight between the two time points. Lack of measurement invariance has been explained as an important threat to the validity of change scores (e.g., Millsap, 2010; Schmitt, 1982).

This thesis reports on both simulation studies and empirical studies with respect to the applicability and the efficiency of CTT and IRT approaches to assessing statistical and clinical change in individual patients. In particular, the following research questions are addressed:

1. Which estimation method based on IRT is the most accurate for detecting individual change as defined by the JT method? (Chapter 2)
2. Are there differences between CTT and IRT with respect to detecting individual change? We answer this question with a simulation study. (Chapter 3)
3. How to apply IRT methods to individual change assessment with real clinical data and to what extent do theoretical differences obtained in Chapter 3 replicate in real data? (Chapter 4)
4. Is there evidence of temporal (i.e., longitudinal) measurement invariance in the Dutch OQ-45? If so, what are its consequences for practical change assessment? (Chapter 5)

Answers to these questions can be informative regarding the conditions under which IRT outperforms CTT and vice versa. Knowledge of strengths and weaknesses of each theory can help practitioners make informed choices when applying them in practice. In addition, investigating issues related to parameter estimation and the operationalization of individual change assessment, in particular of clinical significance, within the context of IRT are also important for implementing IRT-based change assessment in practice. Ultimately, this is a step towards the improvement of individual change assessment.

Overview of the thesis

In chapter 2, we defined and operationalized JT's statistical significance of change within an IRT framework. Using simulated data, we compared three widely-used IRT estimation methods, which are maximum likelihood (ML), weighted maximum likelihood (WML) and the expected a posteriori (EAP) estimation with respect to (1) bias in estimating change scores and their standard errors; and (2) their ability to correctly detect or reject change as defined by the JT method. The three estimation methods were compared for different conditions of test length (i.e., short, long, et cetera) and the magnitude of true change (i.e., small change, large change, et cetera).

Chapters 3 and 4 are dedicated to comparing CTT and IRT with respect to individual change assessment. In Chapter 3, we used a simulation study to compare CTT and IRT with respect to correct (i.e., power) and incorrect (i.e., Type I errors) detection of individual change obtained by means of the JT method. Similar to the previous study discussed in Chapter 2, in this study design factors such as test length and magnitude of true change were manipulated. In Chapter 4, we used real data to present the results of a comparison of CTT and IRT with respect to individual change assessment. In addition to the JT method, we used another change assessment method based on the concept of minimal clinically important difference (MCID). For this study, in a secondary data analysis we used data collected using the Dutch OQ-45 at three mental care institutions in the Netherlands.

In Chapter 5, we examined both the extent to which the assumption of measurement invariance over time is tenable in the Dutch OQ-45 and the consequences of possible violations of measurement invariance for practical change assessment. We examined the stability of the factorial structure from pretest to posttest as well as the possible changes in the way response options are interpreted at these time points. In this study, we used both CTT and IRT methodology to answer the research question.

Chapter 2

Comparison of Three Latent Variable Estimation Methods in Reliable Change Assessment

Abstract

In clinical psychology, it is a common practice to assess the effectiveness of psychotherapy for individual patients. Jacobson and Truax (1991) considered a significance test for individual change scores as an essential part of this assessment and proposed the reliable change index for this purpose. Effective use of the reliable change index requires accurate estimates of the change score and standard error. In this study, we examined three versions of the reliable change index in an IRT context, each version using a different estimate of the latent variable: maximum likelihood, weighted maximum likelihood and expected a posteriori. Using simulated data, for each estimation method we computed the bias and its impact on Type I error rate and sensitivity for detecting reliable change. Results showed that for shorter tests (at most 10 items) reliable change assessment using weighted maximum likelihood produced the smallest bias, but the differences with the other methods were small. For longer tests (at least 20 items), all three reliability change indices performed equally well. We recommend weighted maximum likelihood estimation for short tests and expected a posteriori estimation for long tests.

Chapter 2

2.1 Introduction

Clinical psychologists are often interested in the effectiveness of therapy at the level of an individual patient. In clinical practice, clinicians are advised to regularly assess change in individual patients' mental health to evaluate the outcomes of the treatment they receive. Monitoring of patients provides valuable feedback to both the patient and the therapist, which allows a better fit between an individual's demand for care and the treatment. Research has shown that this approach improves treatment outcomes considerably (Kluger & DeNisi, 1996). Individual-change assessment is also important in experimental studies on the effectiveness of psychotherapy. These studies have traditionally focused on group-mean comparisons (e.g., Kazdin & Wilson, 1978; Meltzoff & Cornreich, 1970). However, according to Jacobson, Follette, and Revenstorf (1984), outcome assessment based on group mean comparisons reveal little or no information about the variability of possible change brought about by a therapy at the individual level (Jacobson et al., 1984).

In order to assess individual change, two questions are in order. The first question is whether observed change reflects real change or mere measurement error. It is well-known that psychological scales are prone to measurement error that can induce random fluctuations in scores over time. When observed change is larger than expected based on random error fluctuations alone, *statistically significant* change is inferred. The second part of individual-change assessment reflects a more substantive issue, which concerns the *clinical significance* of change scores. Small change, even if statistically significant, may reflect change that has little or no practical relevance with respect to the problem from which a patient is suffering. Different methodologies have been proposed to quantify clinical significance. For example, Jacobson and Truax (1991) consider change clinically significant when a patient's score moves from the dysfunctional range at pretest into the functional range at posttest.

The foregoing discussion emphasizes the importance of statistical significance of change as a prerequisite for the assessment of individual change. Without statistical significance, one cannot establish whether change, if any, is real or simply caused by measurement error. In the Jacobson and Truax (JT) model, statistical significance of individual change scores is evaluated using the *reliable change index* (RCI; Jacobson & Truax, 1991, p. 14). The RCI statistic is defined in the context of classical test theory (CTT; Lord & Novick, 1968) and is computed as follows:

$$RCI = \frac{x_2 - x_1}{S_{\text{diff}}},$$

where x_1 and x_2 represent a patients' observed total scores before and after therapy, respectively. S_{diff} is the standard error (SE) of the difference between the two test scores. Thus, RCI expresses the standardized change score. It is assumed to be standard normally distributed in the absence of change.

Using a simple adaptation, the RCI can also be used in the context of item response theory (IRT; e.g., Embretson & Reise, 2000). An IRT approach to assessing reliable change may have some important advantages over CTT methods (e.g., Prieler, 2007). One important advantage is that IRT allows one to use a different SE for each individual depending on his or her location on the latent variable scale. The CTT approach uses a common SE for all individuals, probably underestimating the standard error in the tails of the test-score distribution and overestimating them in the middle. This results in over- or underestimated standardized change. Another important advantage of IRT over CTT is that IRT models describe item characteristics independent of a person population and provide comparable person measurements using different sets of items (Embretson & Reise, 2000). This property is particularly useful, for example, for detecting item bias, adaptive testing, and deriving comparable scores from different clinical scales measuring the same attribute (e.g., Reise, 2005; Reise & Waller, 2009). Other reasons for preferring IRT to CTT are beyond the scope of this paper (for a discussion, see Prieler, 2007).

Assessing reliable change in the context of IRT for each individual requires estimates of the latent variable values at pretest and posttest and their SEs under the postulated IRT model. Hence, it is important that both the latent variable value and the SE are accurately estimated. In practice, three estimation methods are commonly used: maximum likelihood (ML), weighted maximum likelihood (WML), and expected a posteriori (EAP). WML is a modified version of ML, and was designed to reduce bias in the latent variable estimates. EAP is a Bayesian estimation method, which has favorable features compared to other methods within the Bayesian framework (e.g., maximum a posteriori, abbreviated MAP; Embretson & Reise, 2000). For example, compared to MAP, EAP is non-iterative and therefore computationally faster. However, ML, WML and EAP methods can produce biased estimates of either the latent variable, their SEs, or both (Embretson & Reise, 2000; Lord, 1983a; Wang & Wang, 2001). Consequently, since RCI is based on the estimated latent variable values

before and after therapy and the SE of the difference of these values, the potential bias in these estimates may affect the RCI and, as a consequence, deteriorate detection or rejection of reliable change.

The aim of this study was to investigate possible effects of bias in IRT-based RCI indices on the assessment of reliable change using ML, WML and EAP estimation methods. More specifically, we aimed at answering the following two questions:

(1) Which of the three estimation methods (i.e., ML, WML and EAP) produces the smallest bias and is the most efficient (i.e., produces the smallest SE) in estimating change scores and their SEs?

(2) Do ML, WML and EAP produce different Type I error rates and sensitivity in detecting reliable change? Type I error rate is the proportion of patients who are incorrectly classified as having shown a reliable change. Sensitivity is the proportion of patients who are correctly classified as having shown reliable change.

We did a simulation study to answer the research questions. In a simulation study, true latent variable values are known and allow the researcher to assess the discrepancy that bias produces in estimated change scores and their SEs. This article is organized as follows. First, we explain the concept of reliable change in the context of IRT. Second, we discuss the details of the methods used and the results. Third, we discuss the implications for IRT-based assessment of reliable change.

2.1.1 IRT-Based Assessment of Reliable Change

Let θ_{pre} be the true latent variable value of an individual at pretest and let θ_{post} be the true value at posttest. The estimated values are denoted by $\hat{\theta}_{\mathrm{pre}}$ and $\hat{\theta}_{\mathrm{post}}$, respectively. Likewise, let $\sigma_{\hat{\theta}_{\mathrm{pre}}}$ be the true standard errors for pretest measurements and $\sigma_{\hat{\theta}_{\mathrm{post}}}$ for posttest measurements, and let $\hat{\sigma}_{\hat{\theta}_{\mathrm{pre}}}$ and $\hat{\sigma}_{\hat{\theta}_{\mathrm{post}}}$ their estimated values, respectively. The true standard errors can only be obtained if we actually would retest a person infinitely many times under similar conditions. Assuming local independence, in IRT the RCI is defined as

$$RCI_{\hat{\theta}_{\mathrm{post}},\hat{\theta}_{\mathrm{pre}}} = \frac{\hat{\theta}_{\mathrm{post}} - \hat{\theta}_{\mathrm{pre}}}{\sqrt{\hat{\sigma}^2_{\hat{\theta}_{\mathrm{post}}} + \hat{\sigma}^2_{\hat{\theta}_{\mathrm{pre}}}}}.$$

The RCI is assumed to be standard normally distributed. Therefore, absolute values of RCI in excess of a critical z-score corresponding to a desired significance level are considered

reliable. For example, using a two-tailed 10% significance level, $|RCI| \geq 1.645$ indicates reliable change, which can either reflect an improvement or a deterioration of the patient's clinical condition.

2.2 Method

Data Generation

Item-score vectors were simulated using the graded response model (GRM; Embretson & Reise, 2000, pp. 97-102; Samejima, 1969). Let J be the number of items ($j = 1, \ldots, J$). Without loss of generality, the number of item scores is assumed to be the same for each item and equals $M + 1$. Furthermore, let X_j be the random item-score variable with realization x_j ($x_j = 0, \ldots, M$). The GRM defines the response probabilities for each item *j* by means of *M* cumulative response functions, which are defined as

$$P^*_{jx_j}(\theta) = P\left(X_j \geq x_j \middle| \theta\right) = \frac{\exp\left[a_j\left(\theta - b_{jx_j}\right)\right]}{1+\exp\left[a_j\left(\theta - b_{jx_j}\right)\right]} \quad (x_j = 1, \ldots, M) \qquad (1)$$

[$P^*_{j0}(\theta) = 1$ by definition]. In Equation 1, parameter a_j ($a_j > 0$) is the slope parameter and parameter b_{jx_j} is the threshold parameter indicating the value of θ where $P^*_{jx_j}(\theta)$ = .50. Hence, each item is modeled by M threshold parameters b_{jx_j} ($x_j = 1, \ldots, M$). Furthermore, for each item the M threshold parameters have a fixed ordering, $b_{jx_j} \leq \cdots \leq b_{jM}$. The probability of scoring x_j on item *j* can be obtained from Equation 1 using

$$P\left(X_j = x_j \middle| \theta\right) = P^*_{jx_j}(\theta) - P^*_{j(x_j+1)}(\theta)$$

We assumed a standard normal distribution for θ (Embretson & Reise, 2000).

Independent Variables

Estimation methods. The three methods are discussed next in greater detail.

1. ***Maximum Likelihood.*** Let $L(\theta; \mathbf{x}, \boldsymbol{\xi})$ be the likelihood function given an observed item-score vector **x** under the GRM that is defined by the item parameters collected in matrix $\boldsymbol{\xi}$. The ML estimate, denoted $\hat{\theta}_{\mathrm{ML}}$, is the θ value for which the observed item-score vectors is most likely, given the postulated IRT model; that is, the θ value for which $L(\theta; \mathbf{x}, \boldsymbol{\xi})$ reaches its maximum. For item-score vectors that contain only minimum scores 0 or maximum scores M, no finite estimate of θ exists because the likelihood function is either monotonically increasing or decreasing and thus has no maximum. Let $\sigma^2_{\hat{\theta}}$ be the true SE of the estimate, which is the variance of the $\hat{\theta}$s would the same person be measured infinitely many times

under identical conditions. In practice, $\sigma^2_{\hat{\theta}}$ is unknown and has to be estimated under the postulated IRT model. The estimated SE of $\hat{\theta}_{ML}$ is obtained from the information function, denoted $I(\theta)$; that is,

$$\hat{\sigma}^2_{\hat{\theta}_{\mathrm{ML}}} = I\left(\hat{\theta}_{\mathrm{ML}}\right)^{-1} \tag{2}$$

(Embretson & Haviland, 2005). Because in practice the true value θ is unknown, SE is obtained using the information value at $\hat{\theta}_{\mathrm{ML}}$ from the observed likelihood function. Equation 2 is asymptotically true when the number of items goes to infinity.

2. ***Weighted Maximum Likelihood.*** ML estimates are biased to a certain degree, particularly in the tails of the distribution (Lord, 1983; Samejima, 1998). To reduce bias in latent variable estimates under the GRM, Samejima (1998) proposed WML based on Warm's (1989) weighted maximum likelihood method for dichotomous items. WML takes the expected first order bias in ML estimates, denoted $B(\theta)$, into account when estimating θ. In particular, the WML trait estimate, denoted $\hat{\theta}_{\mathrm{WML}}$, is the θ value that maximizes the likelihood function

$$L^*(\theta; \boldsymbol{x}, \boldsymbol{\xi}) = L(\theta; \boldsymbol{x}, \boldsymbol{\xi}) - I(\theta)\, B(\theta).$$

Simulation studies investigating the properties of $\hat{\theta}_{\mathrm{WML}}$ suggested good statistical properties for WML (Wang & Wang, 2001). WML estimates exist also for item-score vectors containing only 0s or maximum scores *M*. Standard errors of WML estimates are obtained using the information function derived from $L^*(\theta; \boldsymbol{x}, \boldsymbol{\xi})$ evaluated at $\hat{\theta}_{\mathrm{WML}}$; that is,

$$\hat{\sigma}^2_{\hat{\theta}_{\mathrm{WML}}} = I^*\left(\hat{\theta}_{\mathrm{ML}}\right)^{-1}. \tag{3}$$

3. ***Expected a Posteriori Estimation.*** EAP is a Bayesian estimation method that combines the likelihood function for the observed item-score vector with a prior distribution of θ representing the assumed population distribution. Let $g(\theta)$ be the prior distribution, which usually is the standard normal (Embretson & Reise, 2000, *p*. 172). The EAP estimate ($\hat{\theta}_{EAP}$) is the expected value of the posterior distribution; that is

$$\hat{\theta}_{EAP} = \frac{\int_{-\infty}^{\infty} \theta L(\theta; \boldsymbol{x}, \boldsymbol{\xi}) g(\theta) d\theta}{\int_{-\infty}^{\infty} L(\theta; \boldsymbol{x}, \boldsymbol{\xi}) g(\theta) d\theta}. \tag{4}$$

The SE of $\hat{\theta}_{EAP}$ equals the SE of the posterior distribution; that is,

$$SE(\hat{\theta}_{\mathrm{EAP}}) = \sqrt{\frac{\int_{-\infty}^{\infty} (\theta - \hat{\theta}_{\mathrm{EAP}})^2 L(\theta; \boldsymbol{x}, \boldsymbol{\xi}) g(\theta) d\theta}{\int_{-\infty}^{\infty} L(\theta; \boldsymbol{x}, \boldsymbol{\xi}) g(\theta) d\theta}}. \tag{5}$$

The integrals in equations 4 and 5 can be approximated by means of numerical integration using a limited number of quadrature points. EAP estimates also exist for item-score vectors containing only minimum scores 0 or maximum scores M. Another advantage of EAP is that it is non-iterative and therefore computationally faster than ML and WML. However, for short tests $\hat{\theta}_{\mathrm{EAP}}$ values are pulled towards the mean of the prior distribution, a bias phenomenon known as shrinkage. Moreover, the shorter the test and the lower the item discriminations are, the larger the effect of shrinkage is on the $\hat{\theta}_{\mathrm{EAP}}$s.

Item parameters. Consistent with the literature on scale properties of clinical and psychological tests (for a review, see Reise & Waller, 2009), we included two conditions for the item parameters in the simulation design. The first condition mimics tests typically employed in clinical settings that measure narrow, unidimensional attributes such as, for example, depression. Because the scales often involve items referring to specific symptomatology (e.g., "I can't sleep well"), these tests typically consist of items with high discrimination power and threshold parameters concentrated at the higher end of the θ scale, the range where pathological patients are located. Therefore, threshold parameters for the first condition were concentrated at the upper half of the θ scale. More specifically, bs were chosen as follows. For each item *j*, the first location parameter b_{j1} was randomly sampled from a uniform distribution defined on $[0; 1]$. Each subsequent location parameter b_{jm} $(m = 2, \dots, M)$ was obtained by adding to the value of $b_{j(m-1)}$ a number sampled from the uniform distribution defined on $[.75;\ 1.25]$. The discrimination parameters (i.e., the a parameters) were sampled from a uniform distribution defined on $[2; 3.5]$. These are typical values for clinical scales (Reise & Waller, 2009).

In the second condition, we simulated data under conditions that mimic tests measuring broader attributes by means of items which have weaker discrimination power (Reise & Waller, 2009). Item parameters were chosen as follows. The bs were spread evenly along the whole θ scale. More specifically, b_{j1}s $(j = 1, \dots, J)$ were sampled from a uniform distribution defined on $[-3; -1]$ and each subsequent b_{jm} $(m = 2, \dots, M)$ was obtained by adding to the value of $b_{j(m-1)}$ a number sampled from a uniform distribution defined on $[1; 1.5]$. The discrimination parameters in this condition were sampled from a uniform distribution defined on $[1; 2.5]$.

Test length. Based on a preliminary literature review (Arthur & Day, 1994; Crowder & Michael, 1991; Gosling, Rentfrow, & Swann, 2003), we used three different test lengths in our study: 5, 10 or 20 items, reflecting typical test lengths used in practice.

Magnitude of change. Following Finkelman, Weiss, and Kim-Kang (2010), true change (denoted δ) was chosen to be either 0 (no change), 0.5 (small change), 1 (medium change) or 1.5 (large change).

The result is a crossed factorial design with 3 (θ estimation method) × 2 (configuration of item parameters) × 3 (test length) × 3 (magnitude of change) cells. In each design cell, we simulated change scores at seven equidistant values of θ_{pre} within the interval between –3 and 3. Change scores were obtained by simulating 1,000 pairs of item-score vectors, each pair containing one item-score vector for θ_{pre} and one for $\theta_{\text{post}} = \theta_{\text{pre}} + \delta$. For each generated item-score vector, we obtained estimates of ϑ and their SEs and computed the change scores and their SEs. For each cell, the result is 1,000 change-score estimates and corresponding SEs. The complete design was replicated 50 times.

Dependent Variables

Bias in IRT change scores and bias in SEs. For each condition, we computed bias in the change scores and bias in the estimated SEs. Let $d(\theta_{\text{pre}}, \theta_{\text{post}})$ be the true change defined by the difference $\theta_{\text{post}} - \theta_{\text{pre}}$ and let $d(\hat{\theta}_{\text{pre}}, \hat{\theta}_{\text{post}})$ be the estimated change that equals $\hat{\theta}_{\text{pre}} - \hat{\theta}_{\text{post}}$. Bias in IRT change scores is defined as:

$$\text{Bias}\left[d\left(\hat{\theta}_{\text{pre}}, \hat{\theta}_{\text{post}}\right)\right] = \overline{d\left(\hat{\theta}_{\text{pre}}, \hat{\theta}_{\text{post}}\right)} - \delta\left(\theta_{\text{pre}}, \theta_{\text{post}}\right),$$

in which $\overline{d(\hat{\theta}_{\text{pre}}, \hat{\theta}_{\text{post}})}$ is the average of the 1,000 simulated change scores.

To compute the bias of the estimated SEs, we first computed the standard deviation of 1,000 replications of $d(\hat{\theta}_{\text{pre}}, \hat{\theta}_{\text{post}})$, which is referred to as *empirical* SE given true change $\delta(\theta_{\text{pre}}, \theta_{\text{post}})$ and denoted by $\sigma[d\left(\hat{\theta}_{\text{pre}}, \hat{\theta}_{\text{post}}\right)|\ \delta(\theta_{\text{pre}}, \theta_{\text{post}})]$. The empirical SE gives the true amount of sampling variation of the estimated change scores if a person would be retested under similar conditions. Bias in the SEs was obtained by taking the differences between the mean of the empirical SEs and the estimated SEs for each of the ML (Equation 2), WML (Equation 3) or EAP (Equation 5) estimation methods; that is,

$$\text{Bias}\left[SE\left(d(\hat{\theta}_{\text{pre}}, \hat{\theta}_{\text{post}})\right)\right] = \overline{\hat{\sigma}[d(\hat{\theta}_{\text{pre}}, \hat{\theta}_{\text{post}})]} - \sigma[d\left(\hat{\theta}_{\text{pre}}, \hat{\theta}_{\text{post}}\right)|\ \delta(\theta_{\text{pre}}, \theta_{\text{post}})].$$

Type I error rates and sensitivity. In each design cell, and for each of the seven equidistant levels of θ_{pre} within each cell, we computed for each θ estimation method the Type I error rate or the sensitivity. The Type 1 error rate is the proportion of persons with δ = 0 showing reliable change due to measurement error. Sensitivity is the proportion of persons with $\delta > 0$ who were correctly identified by RCI as showing true change. Type I error rate and sensitivity were obtained using a two-tailed nominal α level of $.10$; that is, the |RCI| had to exceed 1.645. The choice of α is based on Emons, Sijtsma, and Meijer (2007), who argued that certainty levels of at least .90 are sufficient for making important decisions about individuals.

All computations were done by means of R (R development core team, 2013). To simulate data we used our own code (to be available upon request from the first author). For the estimation of θs and the SEs we used the R-package *catIrt* (Nydick, 2013).

2.3 Results

For the condition representing clinical scales, data simulation for θ_{pre} values equal to -3 and -2 resulted in item-score vectors containing only 0s, rendering ML estimation impossible. Therefore, we present results for this condition only for $\theta_{pre} = -1, 0, 1, 2, 3$.

Bias in change scores. Figures 1 and 2 show bias in estimated change scores. In general, in all conditions of positive change ($\delta > 0$; graphs *b*, c and *d*) WML produced the least biased change scores followed by ML and EAP. The difference between WML and ML was noticeable mainly for the extreme values of θ_{pre} when WML was less biased than ML. WML and ML hardly differed from each other in the middle range of the θ scale. EAP was generally the most biased estimation method. However, in the condition representing clinical scales (Figure 2), the difference between EAP and WML and ML was smaller due to a general decrease of bias in EAP. Moreover, contrary to the condition representing general psychological scales, in the clinical-scale condition EAP produced smaller bias than ML and WML at the lower end of the θ scale ($\theta_{pre} = 0, -1$). In general, all methods produced negative bias and bias was greater for the extreme values of θ_{pre} where test information was the lowest. All other conditions being equal, bias in change scores decreased as the number of items increased. Moreover, increasing the number of items decreased the differences between the bias the three estimation methods produced.

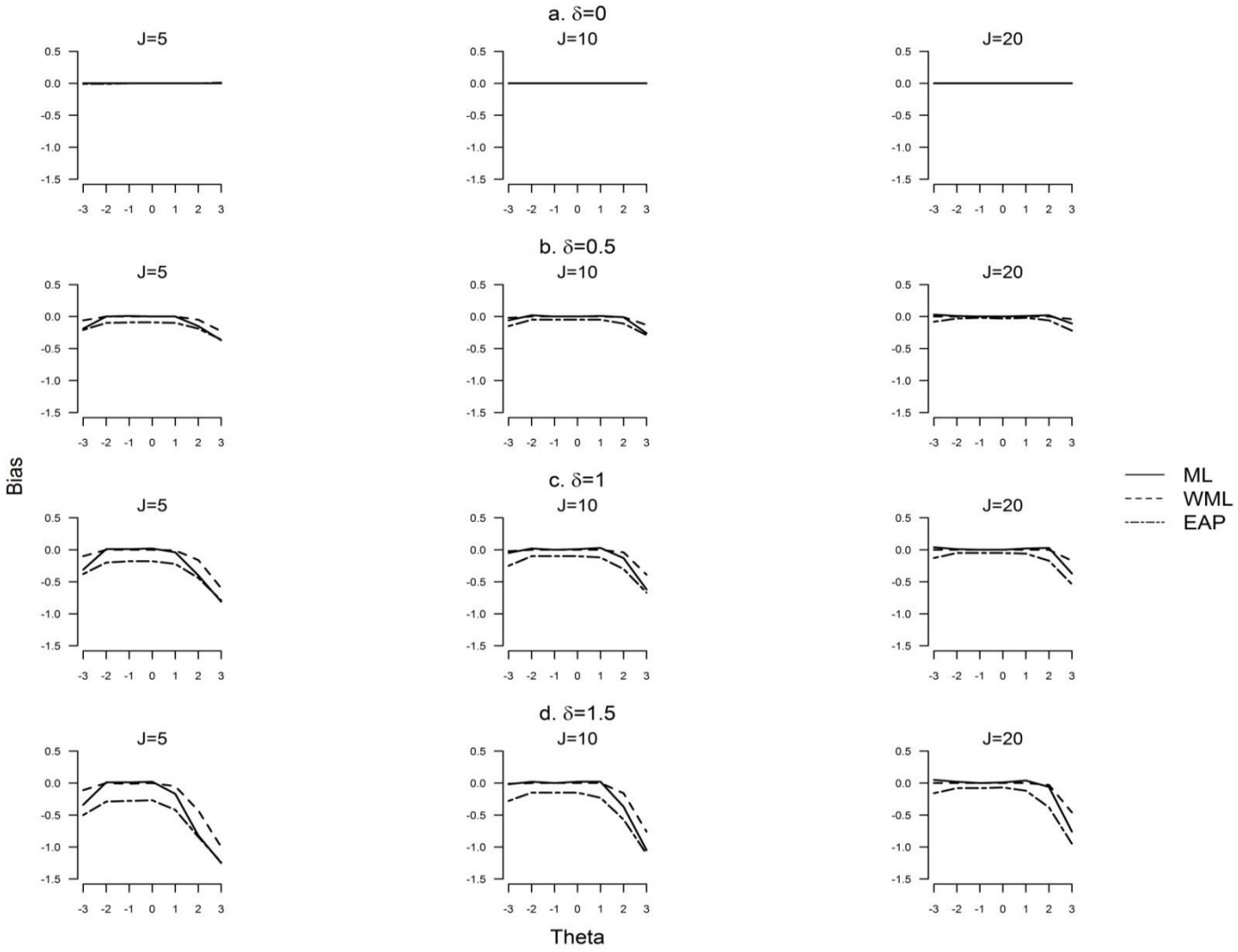


Figure 1. *Bias in estimated change scores for large-range thresholds for four levels of change* (δ) *and three levels of test length (J). The horizontal axis represents the level of* ϑ *at pretest.*

In conditions of no change ($\delta = 0$; figures 1 and 2, graph *a*), bias in change scores was negligible because both $\hat{\theta}_{\text{pre}}$ and $\hat{\theta}_{\text{post}}$ hardly differed from one another.

Bias in the standard errors. Bias in SE was close to 0 in the middle range of the θ scale for all three methods with EAP being slightly more biased than ML and WML (figures 3 and 4). However, in general for extreme θ_{pre} values EAP was less biased than the other two methods. ML and WML differed from each other only at the extremes of the θ scale where ML was less biased than WML.

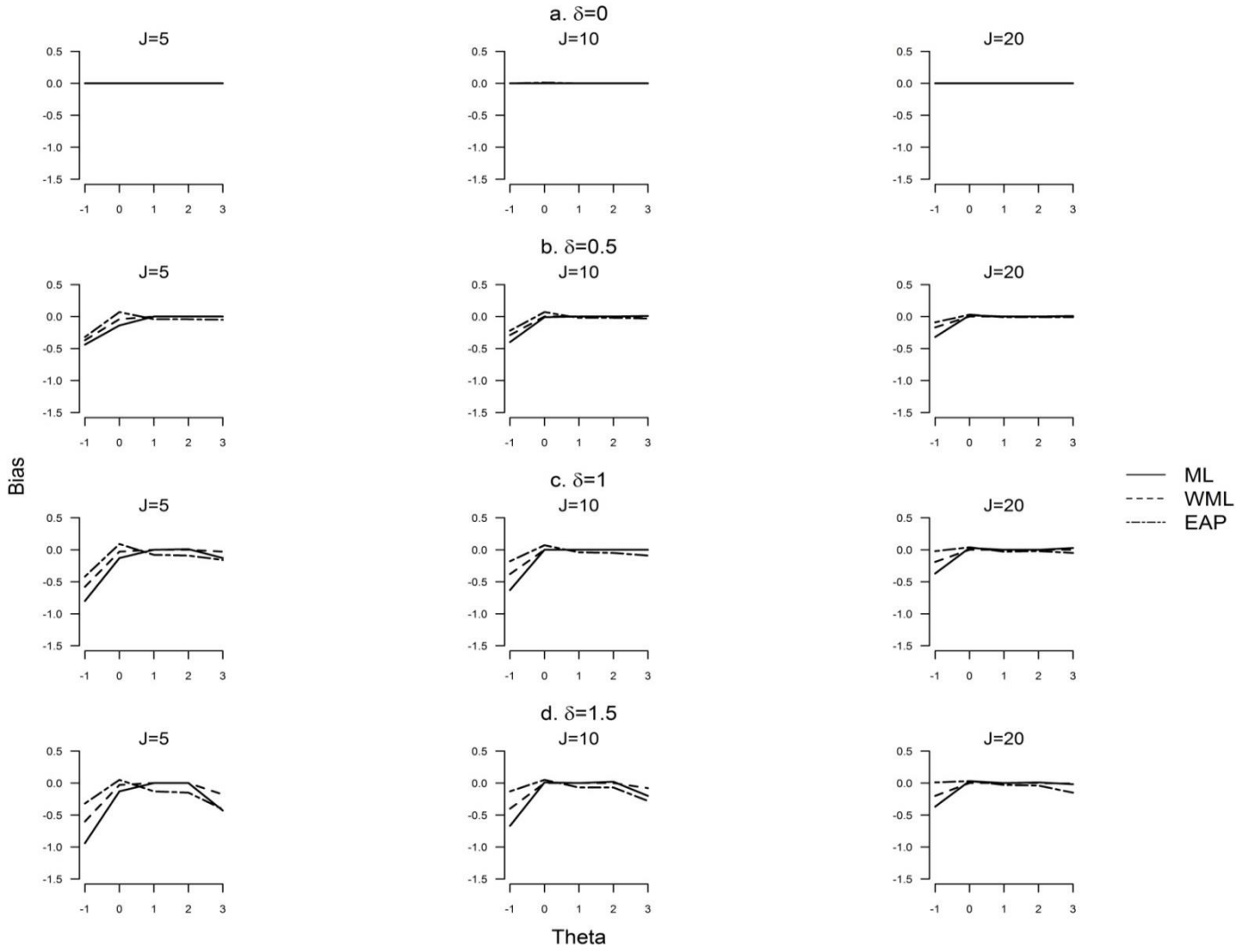


Figure 2. *Bias in estimated change scores for small-range thresholds for four levels of change* (δ) *and three levels of test length (J). The horizontal axis represents the level of* ϑ *at pretest.*

The two methods hardly differed from each other in the middle range of the scale. In general, all methods produced positive bias and bias was greater for the extreme values of θ_{pre} where test information was the lowest. Similar to bias in change scores, bias in SE decreased as the number of items increased. Also, increase of the number of items decreased the differences between the bias the methods ML, WML and EAP produced.

Type I error rates. Table 1 shows Type I error rates for conditions representing psychological (upper panel) and clinical scales (lower panel). Overall, ML and WML produced similar Type I error rates in the middle range of the θ scale. For extreme θ_{pre}s, with ML was

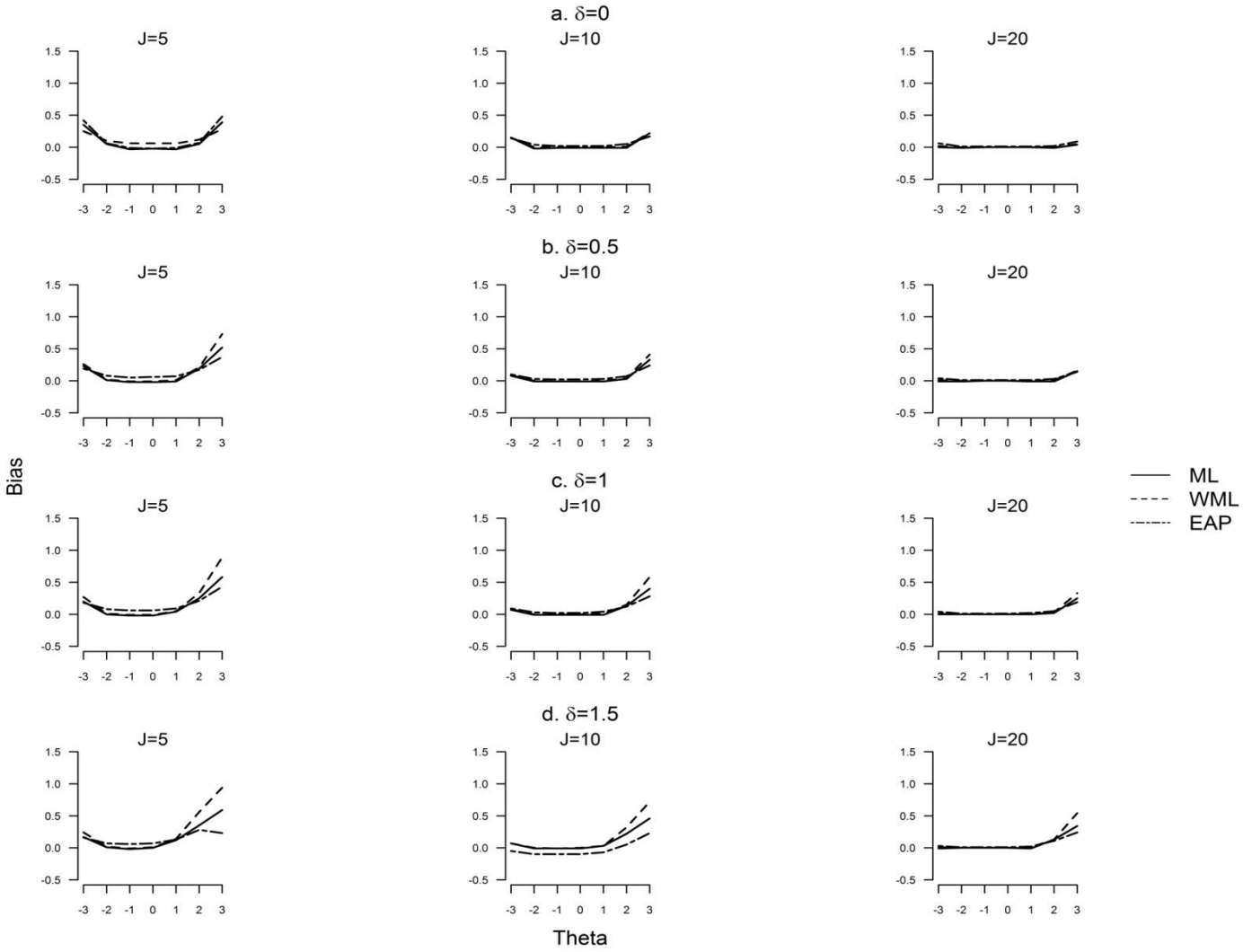


Figure 3. *Bias in estimated SE for large-range thresholds for four levels of change* (δ) *and three levels of test length (J). The horizontal axis in each figure represents the level of* ϑ *at pretest.*

slightly closer to the nominal $\alpha = .10$ level than WML. EAP produced Type I error rates that deviated from the nominal α level more than the other two methods. In general, except for those parts of the θ scale where information was the lowest, Type I error rates for all three estimation methods were generally close to the nominal α level. These results show that under the null hypothesis of no change, methods ML, WML and EAP performed adequately when testing for reliable change in the middle range of the θ scale. For θ_{pre}s where test

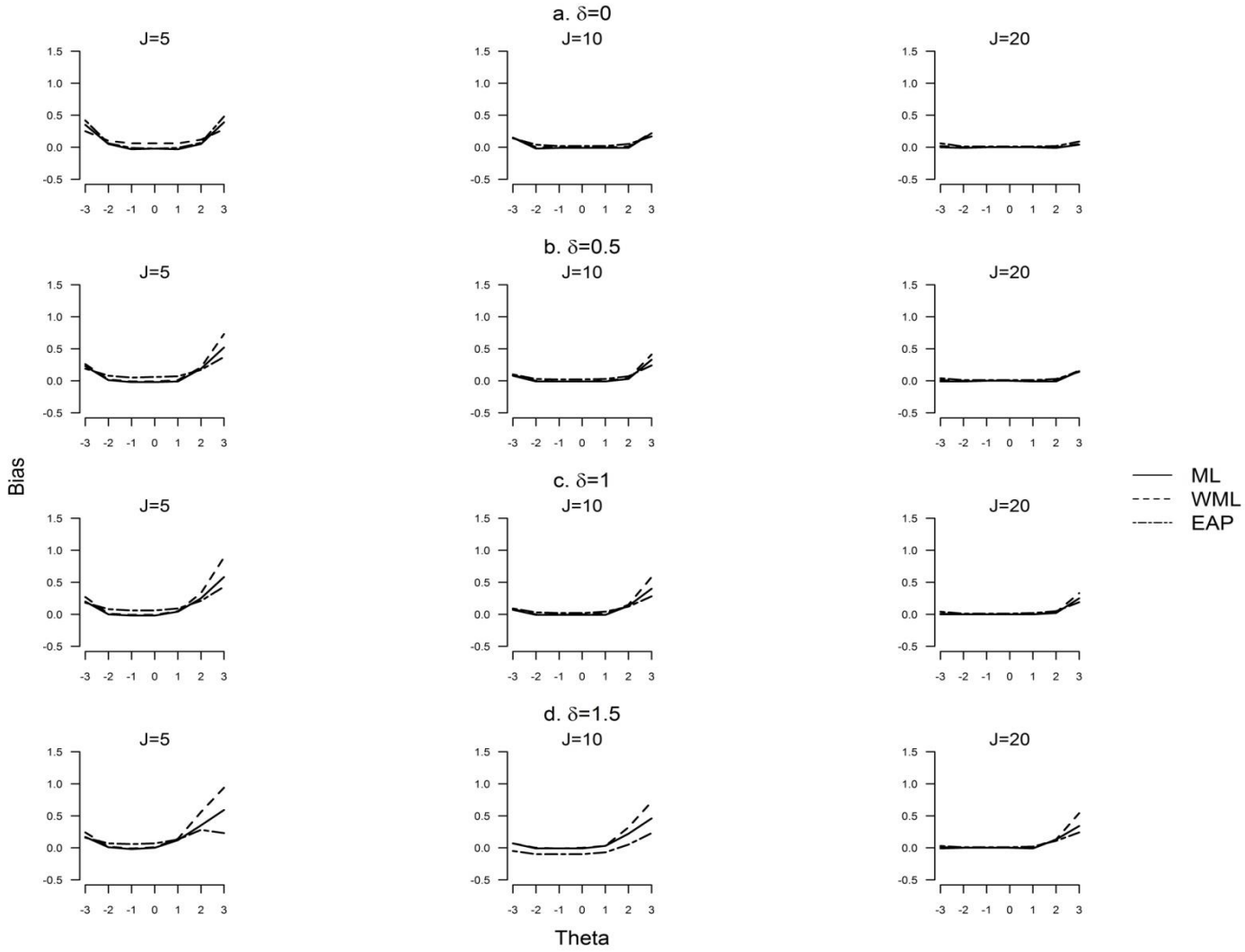


Figure 4. *Bias in estimated SE for small-range thresholds for four levels of change* (δ) *and three levels of test length (J). The horizontal axis in each figure represents the level of* ϑ *at pretest.*

information was the lowest, the empirical Type I error rates the three methods produced were considerably lower than the nominal α level of .10. This means that the RCI is more conservative for detecting reliable change when information is low. Overall, increasing the test length decreased the differences between the Type I error rates and the nominal α.

Sensitivity. In conditions reflecting positive change ($\delta > 0$), in general sensitivity was a little higher for ML and WML than EAP in the middle range of the θ scale (Figure 5, graphs *b*, c and *d*).

Table 1. *Empirical Type I Error Rates (Nominal Significance level (α) = .10) for Seven Latent Variable Values at Pretest, Three Estimation Methods, and Three Test Lengths.*

	Estimation Method								
	ML			WML			EAP		
J =	5	10	20	5	10	20	5	10	20
θ_{pre}									
Large-Range Thresholds (b) and Low Discrimination (a)									
-3	.01	.03	.07	.00	.01	.05	.01	.03	.05
-2	.07	.10	.10	.05	.08	.10	.05	.07	.09
-1	.11	.10	.10	.10	.10	.10	.07	.08	.09
0	.11	.11	.10	.11	.11	.10	.07	.09	.09
1	.11	.10	.10	.10	.10	.10	.07	.08	.09
2	.07	.09	.10	.05	.07	.09	.05	.07	.08
3	.01	.02	.05	.01	.01	.03	.01	.02	.04
Small-Range Thresholds (b) and High Discrimination (a)									
-1	.00[a]	.00	.00	.00	.00	.00	.00	.00	.01
0	.02	.04	.08	.01	.02	.06	.07	.09	.10
1	.10	.10	.10	.09	.10	.10	.09	.10	.14
2	.11	.10	.10	.11	.10	.10	.08	.09	.12
3	.10	.10	.10	.09	.10	.10	.08	.09	.11

Note. ML: Maximum Likelihood; WML: Weighted Maximum Likelihood; EAP: Expected a posteriori; *J* = Number of items. All values are mean Type I error rate across 50 replications, and each replication used data of 1000 simulations. For ML, the number of valid item-score vectors for θ=-3 and 3 ranged between 7 and 764 for large-range and for θ=-1 and 0 between 3 and 532 for small-range threshold parameter conditions.

However, in the condition representing clinical scales, the difference between EAP and the WML and ML methods decreased due to a general increase of EAP's sensitivity (Figure 6, graphs *b*, c and *d*). Moreover, contrary to the previous condition representing general psychological scales, when clinical scales were used, for $\theta_{\mathrm{pre}} = -1$ EAP was more sensitive than ML and WML. In general, all three methods were more sensitive in the range of the θ values where test information was the highest. Sensitivity increased as the number of items

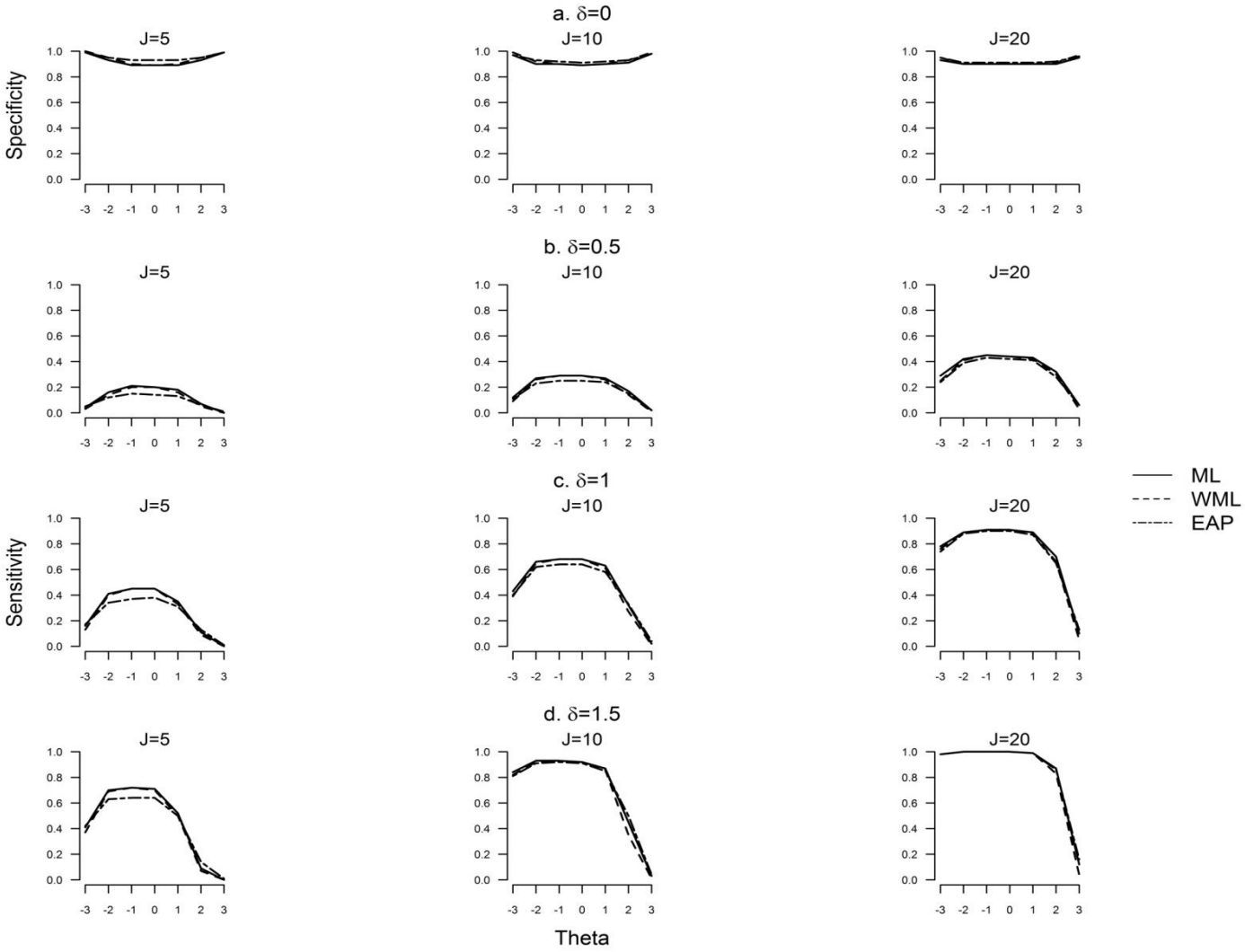


Figure 5. *Detection rates for large-range thresholds for four levels of change* (δ) *and three levels of test length (J). The horizontal axis represents the level of* ϑ *at pretest.*

and the magnitude of change increased. Increasing the number of items also decreased the differences between the sensitivity the three estimation methods produced.

2.4 Discussion

The current study compared ML, WML and EAP estimation methods with respect to reliable change assessment on the θ scale of IRT. Based on our findings we were unable to single out one method that is superior in all aspects, that is, magnitude of bias in change

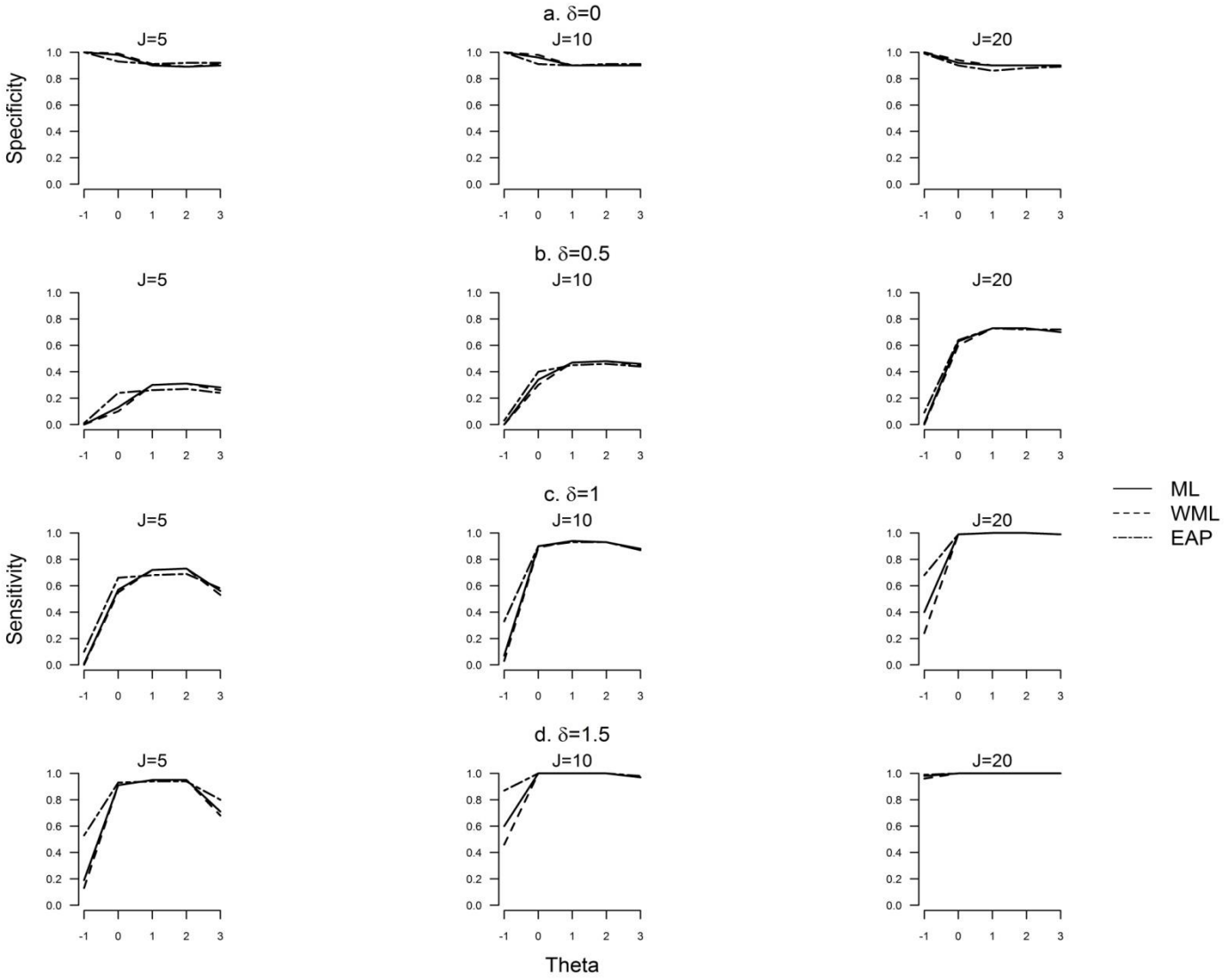


Figure 6. *Detection rates for small-range thresholds for four levels of change* (δ) *and three levels of test length (J). The horizontal axis represents the level of* ϑ *at pretest.*

scores and SEs, Type I error rates and sensitivity for detecting reliable change. Moreover, which reliable change index performs best depends on the combination of multiple factors such as test length and the available local information rather than on the estimation method alone. For example, with respect to sensitivity for reliable change assessment, WML performed better when tests were short and items had low discrimination, whereas EAP and ML performed better when tests were long and items had high discrimination. One of the recommendations for test practitioners is to avoid using short tests with low discrimination

because these tests may produce more biased estimates of change scores and SEs, higher Type I error rates and lower sensitivity for detecting reliable change. Short tests had a higher negative impact on EAP than ML and WML but the difference was small. This result was due to the fact that the smaller the number of items is in a test, the greater the influence of the prior distribution is on the EAP estimates which results in more shrinkage of these estimates towards the mean of the prior.

More shrinkage toward the mean was also found for tests with low item discrimination. However, in the condition representing clinical scales the differences between the methods were negligible irrespective of the number of items. These items had high discrimination, which reduced the effect of the prior on the EAP estimates, which in turn reduced bias of estimated change scores and SEs due to shrinkage.

Our results also show that increasing test length and item discrimination decreased the difference between the sensitivity of the three estimation methods. For tests containing 20 highly-discriminating items (i.e., clinical scales), sensitivity was equal for methods ML, WML and EAP. However, for larger number of items, the estimation process takes more time, and since EAP is characterized by lowest computational burden, it can speed up the estimation process for longer tests. Therefore, when longer tests are used for assessing reliable change we think it is most efficient to use EAP rather than ML and WML. Researchers and practitioners who use short tests containing items with low discrimination are advised to use ML or WML when assessing reliable change because these two methods were less biased and slightly more sensitive than EAP.

Increasing test length and item discrimination improved sensitivity but only for situations that are rare in practice. That is, sensitivity was generally low unless true change was equal to 1 or higher ($\delta \geq 1$). In clinical practice, a true change of such magnitude may be difficult to attain. Clinical practitioners generally accept a true change of 0.5 as the minimal change that bears clinical importance. In this study, maximum sensitivity on average ranged from 0.2 and 0.7 in the conditions where true change equaled 0.5. Similar results were found in the context of CTT (Kruyen, Emons, Sijtsma, 2012) and IRT (Kruyen, Emons, Sijtsma, 2012, 2014). This means that despite the advantages of IRT-based RCIs, simulation studies suggest that IRT methods still have low power to find the minimum change of substantive clinical interest. Future studies may further explore the usefulness of IRT-based methods relative to

Chapter 2

CTT approaches for detecting both reliable and clinically significant change in real clinical applications.

Chapter 3

Comparison of Classical Test Theory and Item Response Theory in Individual Change Assessment

Abstract

Clinical psychologists are advised to assess clinical and statistical significance when assessing change in individual patients. Individual-change assessment can be conducted using either the methodologies of classical test theory (CTT) or item response theory (IRT). Researchers have been optimistic about the possible advantages of using IRT rather than CTT. However, little empirical evidence is available to support the alleged superiority of IRT for individual-change assessment. In this study, we compared CTT and IRT with respect to their Type I error and detection rate. IRT was found to be superior to CTT in individual-change detection for tests consisting of at least 20 items. For shorter tests, compared to IRT, CTT detected change in individuals more often.

Chapter 3

3.1 Introduction

Individual-change assessment plays an important role in clinical practice where clinicians are interested in the effectiveness of treatments for individual patients rather than the average improvement of groups of patients as a whole. The assessment of individual change in clinical contexts can be done using either the methodologies of classical test theory (CTT; e.g., Jacobson & Truax, 1991; Lord & Novick, 1968) or item response theory (IRT; e.g., Embretson & Reise, 2000; Prieler, 2007; Reise & Haviland, 2005). CTT approaches are familiar to most clinicians and therefore widely used, but IRT methods are also gaining popularity.

Several authors have argued that IRT is superior to CTT (e.g., Prieler, 2007; Reise & Haviland, 2005). The most important difference between CTT and IRT is that in CTT one uses one common estimate of measurement precision, which is assumed to be equal for all individuals irrespective of their attribute levels. However, in IRT measurement precision depends on the latent attribute value. As a result, CTT and IRT may differ with respect to their conclusions about statistical significance of change.

There are arguments favoring IRT that are worth mentioning. IRT models, including the popular two-parameter logistic model and the graded response model (Embretson & Reise, 2000), take the pattern of the item scores into account when inferring latent attribute scores, which means that the latent attribute values at pretest and posttest may differ even when the classical pretest sum score and the classical posttest sum score are equal. As a result, IRT may reveal subtle changes in individuals' mental health that would go unnoticed when using the sum scores which ignore the pattern of the scores typical of CTT. Finally, IRT facilitates adaptive testing, which allows researchers to use different questions at pretest and posttest provided that the items are all calibrated on the same scale. A major drawback of IRT approaches to change assessment is their reliance on the availability of accurate estimates of the item parameters and model fit, which may be costly and difficult to realize.

Empirical studies comparing CTT and IRT have shown ambiguous results (e.g., Brouwer, 2013; Sébille et al., 2010), suggesting that the CTT approach may be as effective as IRT for assessing individual change. However, so far a systematic head-to-head comparison of the two approaches in the context of individual-change assessment has not been done. Given the importance of individual-change assessment in clinical settings, the optimism about IRT methods, and the ambiguous empirical comparison results for CTT and IRT, we investigated to what extent CTT and IRT differ with respect to classifying patients into different categories

of individual change based on the combination of clinical and statistical significance. The results of the comparison can help clinicians and researchers make more informed decisions about scoring tests and assessing change.

This article is organized as follows. First, we explain Jacobson and Truax's (Jacobson & Truax, 1991; henceforth JT) operationalization of clinically and statistically significant change in the CTT context and we extend their approach to IRT. Then we discuss the design and the results of a simulation study which compares CTT and IRT with respect to Type I error rate and individual change detection. Finally, we discuss the implications of the results and provide recommendations for researchers and clinicians working in clinical settings.

3.1.1 Operationalization of Individual Change in CTT and IRT

CTT Approach of Jacobson and Truax

Reliable change. Let X be the sum score based on the J items in the test, with item scores denoted by X_j $(j = 1, \dots, J)$, so that $X = \sum_{j=1}^{J} X_j$. Let X_{pre} and X_{post} be the sum scores on the pretest and the posttest, respectively, briefly called pretest and posttest scores. In what follows, we assume that pretest and posttest scores are obtained on identical tests or questionnaires. Statistical significance of change is assessed by means of the reliable change index (RCI), which JT (1991) defined as follows. Let $d = X_{\text{post}} - X_{\text{pre}}$ be the change score for an individual patient. Assuming that higher scores reflect worse health conditions, $d < 0$ suggests improvement and $d > 0$ suggests deterioration. Furthermore, let SEM_d be the standard error of measurement (SEM) of change score d. To assess individual change, the following assumptions are made: (*a*) equal measurement precision at pretest and posttest, that is, $SEM_{X_{\text{pre}}} = SEM_{X_{\text{post}}} = SEM_X$; (*b*) uncorrelated measurement errors between pretest and posttest; and (*c*) measurement invariance, that is, the test is measuring the same latent attribute at pretest and posttest and the answer categories are interpreted in the same way at pretest and posttest. Using these assumptions, we obtain $SEM_d = \sqrt{2} \times SEM_{X_{\text{pre}}}$. JT (1991) defined the RCI as

$$RCI_{CTT} = \frac{d}{SEM_d}. \tag{1}$$

The RCI is assumed to be standard normally distributed in the absence of change. An RCI with an absolute value that exceeds the critical z-score corresponding to a desired significance

level is considered to represent reliable change. For example, at two-tailed significance level of $.10$, $|RCI_{CTT}| \geq 1.645$ indicates reliable change, which can either mean improvement or deterioration.

Clinical significance assessment. JT assessed clinical significance by evaluating whether a patient's pretest score moved from the dysfunctional score range to the functional score range at posttest; JT defined these ranges in three ways. Let X_{cut} denote the clinical cutoff separating functional and dysfunctional score ranges. Because we assume that higher scores reflect worse clinical conditions, clinical significance is inferred if $X_{\text{pre}} > X_{\text{cut}}$ and $X_{\text{post}} < X_{\text{cut}}$. JT defined functional and dysfunctional score ranges based on cut scores from either the distributions of the scores in the functional or healthy population, the dysfunctional or clinical population, or both. They proposed to use one of the following cutoffs: (a) the 90^{th} percentile of the score distribution in the functional population; (b) the 10^{th} percentile of the score distribution in the dysfunctional population; or (c) the average of the means of the score distributions in the functional and the dysfunctional populations. JT advocated the use of cutoff (c), but this cutoff requires data sampled from both a functional and dysfunctional populations, and such datasets are often unavailable. For a more elaborate discussion of the pros and cons of different cutoffs, see JT (1991; also, Jacobson, Roberts, Berns, & McGlinchey, 1999; and the explanation and Figure $A1$ in the appendix to this chapter).

Based on the combination of clinical and statistical significance of change scores, and the direction of the observed change, patients can be classified into one of five exhaustive and mutually exclusive change categories (e.g., Bauer, Lambert, & Nielsen, 2004), labeled (i) *no change*; that is, change is neither statistically nor clinically significant; (ii) *improvement*; that is, change indicates better functioning and is statistically but not clinically significant; (iii) *recovery*; that is, change indicates better functioning which is both statistically and clinically significant; (iv) *deterioration;* that is, change indicates worse functioning which is statistically but not clinically significant; and (v) *clinically significant deterioration;* that is, change indicates worse functioning which is both statistically and clinically significant. Two remarks are in order. First, for persons to be classified as having deteriorated to a clinically significant degree, change has to be statistically significant and the pretest and posttest scores have to belong to the functional and dysfunctional ranges, respectively. Second, *no change* means that the observed change is too small to be statistically significant. In practice, non-significant change means that more information is needed before reliable conclusions

about individual change are made. Thus, one should not conclude that no change has occurred.

IRT Perspective

Reliable change. The assessment of statistical and clinical significance of individual change in the context of CTT can be readily extended to IRT. Let $\hat{\theta}_{\text{pre}}$ and $\hat{\theta}_{\text{post}}$ be the estimated latent attribute values at pretest and posttest under the postulated IRT model, respectively. Furthermore, let $SE(\hat{\theta}_{pre})$ and $SE(\hat{\theta}_{post})$ be the standard errors for the estimated pretest and posttest scores, respectively. Assuming independent observations at the individual level, the RCI in the context of IRT is defined as

$$RCI_{\text{IRT}} = \frac{\hat{\theta}_{\text{post}} - \hat{\theta}_{\text{pre}}}{\sqrt{SE\left(\hat{\theta}_{\text{pre}}\right)^2 + SE\left(\hat{\theta}_{\text{post}}\right)^2}}. \quad (2)$$

Equation (2) requires estimates of the latent attribute values, $\hat{\theta}_{\text{pre}}$ and $\hat{\theta}_{\text{post}}$. Research showed that weighted maximum likelihood (WML) produces estimates having the smallest bias and the greatest precision (e.g., Jabrayilov, Emons, & Sijtsma, 2014; Wang & Wang, 2001). Standard errors are obtained by means of the information function (e.g., Reise & Haviland, 2005). IRT-based individual-change assessment requires the availability of accurate estimates of all item parameters, for example, by means of multiple-group IRT models when data are obtained from both general and clinical populations (e.g., Jabrayilov, Emons, De Jong, & Sijtsma, 2015). Henceforth, we assume that this requirement is met and use the true parameters for estimating the person parameters. In addition, unlike CTT, IRT methods do not require pretest and posttest measurements to be based on the same items as long as all items are calibrated on the same scale. However, to fairly compare CTT and IRT, we used the same items at pretest and posttest.

Clinical-significance assessment. In an IRT context, clinical significance can be assessed by examining whether the posttest score has passed a clinical cutoff. The crucial difference between CTT and IRT is that in CTT the cutoffs are based on the distribution of the sum scores X, whereas in IRT they are based on the latent θ distribution. For example, in the IRT context JTs cutoff (a) would be the 90^{th} percentile of the θ distribution in the functional population.

3.1.2 Comparing Measurement Precision in CTT and IRT

One of the main arguments for favoring IRT methods is that they allow using the local precision of the estimated scores, $SE(\hat{\theta})$, to test change for significance, whereas in CTT one common population-level SEM is used for all persons. Because the population-level SEM used in CTT is the average of the individual SEMs (Mellenbergh, 2011, p. 119) which vary across individuals, using the SEM results in overestimating measurement precision of the scores in the tails of the distribution and underestimating it in the middle of the distribution (e.g., Mollenkopf, 1949); see Figure 1 (upper graph) for population-level constant SEM and the empirical standard error for observed scores. Therefore, using SEM may bias decisions based on RCI.

The standard errors in Equation (2) are usually obtained using the Fisher information function evaluated at $\hat{\theta}$, but are only accurate if the number of items is sufficiently large, say, more than 20 (Magis, 2014). Clinical practice shows a tendency for using short scales in order to minimize the burden on patients (Emons, Sijtsma, & Meijer, 2007; Kruyen, Emons, & Sijtsma, 2013a, b; 2014). When θ is estimated from a limited number of discrete item scores, asymptotic results no longer apply and the corresponding estimated standard errors may be inaccurate (Jabrayilov et al., 2014). To illustrate this point, suppose that one repeatedly tests the same patient having an extremely high θ value under identical conditions. Hypothetically, one expects the same pattern of J maximum item scores at each replication; hence, the patient obtains the same $\hat{\theta}$ each time and the empirical standard error is small. For high θ values, however, test information is low and thus the asymptotic standard error is large. Hence, for extreme θ values IRT methods tend to overestimate the empirical standard errors when scales are short. For a 10-item test with varying difficulties, Figure 1 (lower graph) shows the relationship between estimated asymptotic standard errors and empirical standard errors.

Given the differences between CTT and IRT with respect to change assessment, depending on which method one uses, we expect that different conclusions may be drawn about change in individual patients. Because of the high expectations regarding IRT as a refinement and an improvement relative to CTT, we were particularly interested in finding out whether, compared to CTT, IRT produces more precise Type I error and higher detection rates of clinically significant change. We used a simulation study to this end.

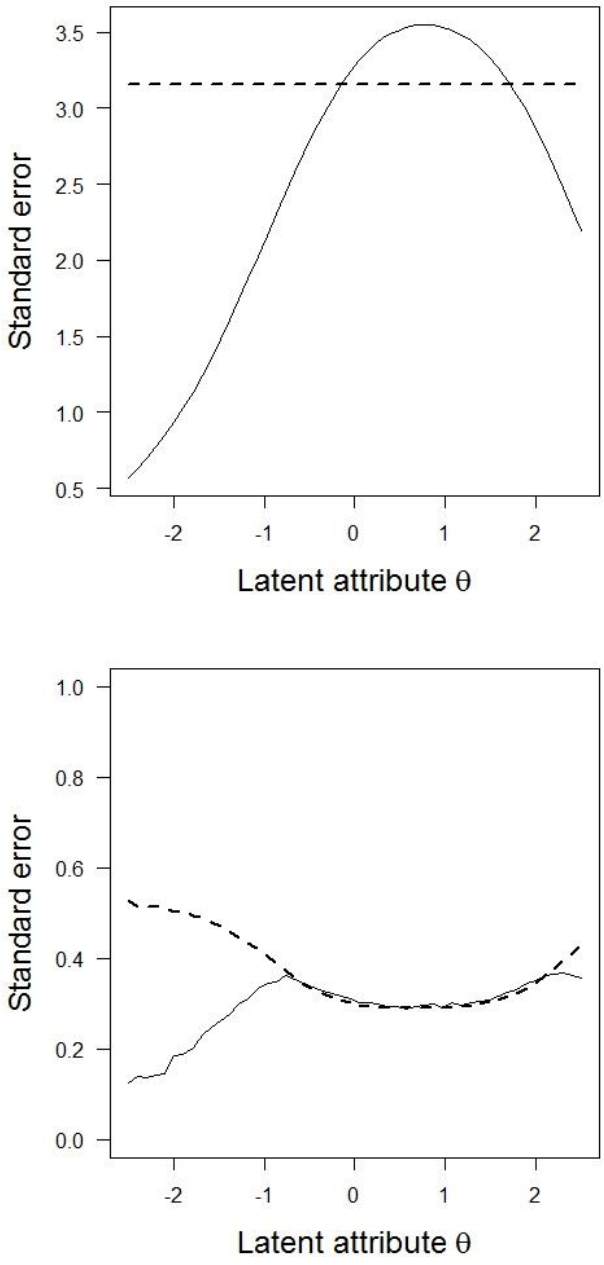


Figure 1. *Comparison of standard errors of estimated person scores in CTT (upper graph) and IRT (lower graph). Note. Upper panel: solid line represents empirical standard error of X as a function of latent attribute θ, dashed line represents SEM. Lower panel: solid line represents empirical standard error of WML estimates $\hat{\theta}$ as a function of θ, dashed line represents the asymptotic standard error based on square root of the inverse of Fisher's information function.*

3.2 Method

Data Generation

Person characteristics. In both the healthy and clinical populations, we assumed normal distributions for latent attribute θ with variance of 1 and means of 0 and 0.5, respectively. Because within-population variances equaled 1, using Cohen's d (Cohen, 1988, p. 26) the difference between the means corresponded to a medium effect size between the healthy and clinical populations. Standard normality of θ in the healthy population was an arbitrary choice that serves to identify the θ scale (e.g., Embretson, 2006).

Test and item characteristics. Pretest and posttest item scores were modeled using the graded response model (GRM; Embretson & Reise, 2000, pp. 97-102; Samejima, 1969). We assumed invariant item parameters between pretest and posttest (i.e., measurement invariance). Let $M + 1$ denote the number of ordered item scores for an item, and let item score X_j have realizations x_j $(x_j = 0, \dots, M)$. The GRM models the probabilities of obtaining a particular item score x_j or a higher score by means of M cumulative response functions, each defined by a two-parameter logistic function,

$$P^*_{jx_j}(\theta) = P\left(X_j \geq x_j \middle| \theta\right) = \frac{\exp\left[a_j\left(\theta - b_{jx_j}\right)\right]}{1+\exp\left[a_j\left(\theta - b_{jx_j}\right)\right]}, \quad x_j = 1, \dots, M. \qquad (3)$$

By definition, $P^*_{j0}(\theta) = 1$ and $P^*_{j,M+1}(\theta) = 0$. The probabilities of obtaining score x_j can be obtained by subtracting the cumulative response probabilities, for $X_j \geq x_j$ and $X_j \geq x_j + 1$ (Embretson & Reise, 2000, p. 99). In Equation (3), a $(a_j > 0)$ represents the slope parameter for item j indicating how well the items discriminates between respondents with different levels of θ, and b_{jx_j} is the threshold parameter indicating the value of θ where $P^*_{jx_j}(\theta) = .50$ and the location on the θ scale where the response function has its maximum slope discriminating different θs best. Hence, each item was modeled by M threshold parameters b_{jx_j} $(x_j = 1, \dots, M)$, which had a fixed ordering $b_{j1} \leq \cdots \leq b_{jM}$. In our study, items were scored from 0 to 4, higher scores indicating more distress. Hence, each item had four b parameters $(M = 4)$.

Item parameters were chosen in two conditions. To guarantee a fair comparison between CTT and IRT, we chose the item parameters such that adequate psychometric properties were obtained *both* in terms of CTT and IRT. In the first condition, scale items were similar with respect to difficulty representing narrow attributes such as depression and

anxiety (Reise & Waller, 2009). We sampled discrimination parameters (a) from $U(1.5;\ 2.5)$. Following Emons et al. (2007), the threshold $\mathbf{b}$s were sampled as follows. Let $\bar{b}_j$ represent the average threshold of item j. For each item, we first sampled $\bar{b}$ from $U(0;\ 1.25)$ and then the four individual $\mathbf{b}$s were obtained as follows: $b_{j1} = \bar{b}_j - 1$, $b_{j2} = \bar{b}_j - 0.5$, $b_{j3} = \bar{b}_j + 0.5$, $b_{j4} = \bar{b}_j + 1$; hence, item mean variation was small. The name of *homogeneous item-difficulty condition* suggested that the mean item-level difficulties were concentrated on a limited range of the ϑ scale.

The second condition represented the characteristics of tests that typically measure potentially broader attributes such as personality traits and quality of life. Reise and Waller (2009) argued that the item difficulties in broad-attribute tests are usually spread across the entire latent attribute scale and on average have somewhat lower discrimination than items in narrow-attribute tests. Therefore, compared to the previous condition we sampled the discrimination parameters (as) and the mean thresholds $\bar{b}_j$s from a wider interval, the as from $U(1;\ 2.5)$ and $\bar{b}_j$s from $U(-1.5;\ 2.5)$. The bs were selected such that the expected mean item scores also varied from low to high. This resulted in bs that were located closer to the $\bar{b}_j$s than in the *homogeneous item-difficulty condition*. The bs equaled $b_{j1} = \bar{b}_j - 0.5$, $b_{j2} = \bar{b}_j - 0.2$, $b_{j3} = \bar{b}_j + 0.2$, $b_{j4} = \bar{b}_j + 0.5$. The second condition's name is *heterogeneous item-difficulty condition*, expressing spread of item-level difficulties along the entire latent attribute scale. The healthy and clinical populations had mean coefficient alpha's at least equal to $.7$, and item-rest score correlations which exceeded $.3$. In the *homogeneous item-difficulty condition*, mean item scores ranged from 0.81 to 2.52 (on a scale running from 0 to 4) and in the *heterogeneous item-difficulty condition* from 0.11 to 3.75. Hence, the simulation set-up generated data that are realistic both in terms of CTT and IRT characteristics.

Determination of Cutoffs for Assessing Clinical Significance

Clinical cutoffs in IRT. Following JT (1991), we defined three different cutoffs: that is, for cutoff a we placed the cutoff at the 90^{th} percentile of the θ-distribution in the healthy population (i.e., $\theta_{\text{cut}} = 1.28$), for cutoff b at the 10^{th} percentile of the θ-distribution in the clinical population (i.e., $\theta_{\text{cut}} = -0.78$), and for cutoff c we chose the average of the two population means of θ (i.e., $\theta_{\text{cut}} = 0.25$).

Clinical cutoffs in CTT. Because in CTT the clinical cutoffs are derived from the sum-score (X) distribution, which depends on both the item characteristics and the latent attribute (θ) distribution, we first obtained the population-level *X* distributions given the IRT item and person parameters and then we determined the JT cutoffs a, b and c from these distributions. In particular, let the item parameters of the GRM be collected in matrix $\boldsymbol{\xi}$ of order J by 5 (1 slope and 4 threshold parameters). Furthermore, for the healthy (indexed by H) and clinical populations (indexed by C), let $f_H(X|\boldsymbol{\xi})$ and $f_c(X|\boldsymbol{\xi})$ be the discrete marginal distributions of X given item parameters $\boldsymbol{\xi}$. To obtain the marginal sum-score distributions, in each population the θ-distribution was approximated using 500 quadrature points. For cutoff a, we selected the value of X that was closest to the 90^{th} percentile of $f_H(X|\boldsymbol{\xi})$; for cutoff b, we selected the *X*-value closest to the 10th percentile of $f_C(X|\boldsymbol{\xi})$; and for cutoff c, we used the average of the two means of the two marginal X-distributions. See the online supplement for details.

Simulation Design

The following four design factors were used:

1. **Change-assessment method.** CTT and IRT.
2. **Test length.** In order to mimic scales used in practical clinical contexts, test length was either 5, 10 or 20 items. Examples of tests with similar test lengths are Outcome Questionnaire OQ-45 (Lambert et al., 1996; Social Role subscale: 9 items; Interpersonal Relations subscale: 11 items; and Symptom Distress subscale: 25 items), Montgomery-Asberg Depression Rating Scale (Montgomery & Ashberg, 1979; 10 items), and Beck Depression Inventory (Beck, Ward, Mendelson, Mock, & Erbaugh, 1961; 21 items).
3. **Magnitude of true change.** True change had four levels: $\delta = 0$ (no change), $\delta = -0.5$ (small change), $\delta = -1$ (medium change), and $\delta = -1.5$ (large change). Because clinical treatment focuses on improvement, we concentrated on change reflecting improvement (i.e., $\delta \leq 0$). Also, since in our simulation study the direction of the change does not have an intrinsic meaning, we considered the results also representative of detecting deterioration (also, see Kruyen et al., 2014).
4. **Item characteristics.** Homogeneous and heterogeneous mean item difficulties.

The design was a fully crossed factorial, with 2 (CTT, IRT) $\times$ 3 (Test Length) $\times$ 4 (True Change) $\times$ 2 (Item Characteristics) $=$ 48 cells. In each cell, change scores were simulated as

follows. We chose 500 equally spaced pretest θ values (i.e., θ_{pre}) between -2.5 and 3.5. For each θ_{pre} value, we simulated $5{,}000$ pairs of item-score vectors, one for the pretest and one for the posttest. The θ value used for generating posttest data depended on the pretest value θ_{pre} and true change δ; that is, $\theta_{\text{post}} = \theta_{\text{pre}} + \delta$. For each pair of item-score vectors, we estimated pre- and posttest latent attribute values ($\hat{\theta}$) using WML estimation and computed the observed change and the RCI_{IRT} (Equation 2). For each pair, we also computed the sum scores at pretest and posttest, the observed change (d) and the RCI_{CTT} (Equation 1) using the population-based value of the SEM in the clinical population (see online supplement, Table $A1$, for details). This resulted in $5{,}000$ replications of CTT and IRT-based individual-change assessment at each value of θ_{pre}. The complete design was replicated 100 times, each time using newly sampled α_j and $\bar{\text{b}}_j$ parameters.

Dependent Variable

The dependent variable was the individual classification with respect to individual change in the following three exhaustive and mutually exclusive categories of individual change: (i) no change; (ii) improvement; and (iii) recovery. Based on Emons et al. (2007), we used a $.10$ significance level for testing statistical significance. Emons et al. (2007) argued that for high-stakes decisions certainty levels of .90 or higher are acceptable.

To present the results, we made a distinction between classifications under the zero true-change condition ($\delta = 0$) and the other conditions (i.e., $\delta < 0$). In the zero true-change condition, patients whose observed scores showed recovery or improvement did not really change, and hence constituted Type I errors. Thus, the percentage of classifications in either the recovery or improvement condition (i.e., patients showing reliable change irrespective of whether the change is clinically significant) were reported as Type I error rates. For all other conditions ($\delta < 0$), we reported population-level percentages of correct classifications into either improvement or recovery categories. The population-level percentage is a weighted average of the percentages at all θ levels, where the weights are based on the θ-distributions (see Appendix). Overall percentages were referred to as detection rates.

For the simulations, we developed dedicated software in C++. All other computations were done in R (R development core team, 2014). Source code for C++ and R are available upon request from the corresponding author.

3.3 Results

Zero Change. Table 1 shows the population-level Type I error rates in the zero-change (i.e., $\delta = 0$) condition.

Table 1. *Population-Level Type I Error Rates (Entries are Means Across 100 Replications) for Detecting Reliable Change at Nominal Significance Level of .10, for Varying Test Length and Test Model, and Two Item-Location Spreads.*

Item difficulty	Test length					
	5		10		20	
	CTT	IRT	CTT	IRT	CTT	IRT
Homogeneous	.10	.07	.10	.08	.10	.09
Heterogeneous	.09	.05	.09	.08	.09	.09

Note. Values are means across 100 replications. Standard errors of the means ranged from 0.0005 to 0.001.

In general, CTT Type I error rates were closer to the nominal $\alpha = .10$ than those of IRT. Both in the homogeneous and heterogeneous item-difficulty conditions, CTT had equal Type I error rates irrespective of test length. In contrast, for IRT increasing the test length pulled the Type I error rates closer to the nominal Type I error.

To better understand how the two methods differ with respect to their Type I error rates, we plotted the Type I error rate as a function of latent attribute θ (Figure 2). For CTT, for homogeneous tests Type I error rates were above nominal level α in the middle range of the clinical population distribution and below nominal level α at the tails. However, for heterogeneous tests, Type I error rates were at or below nominal α. For IRT, both in the homogeneous and heterogeneous item-difficulty conditions the Type I error rates were at or below the nominal level across the entire scale range, with larger differences at the extremes. These results are consistent with standard errors being underestimated by the group-based SEM from CTT in the middle range of the distribution and overestimated in its tails.

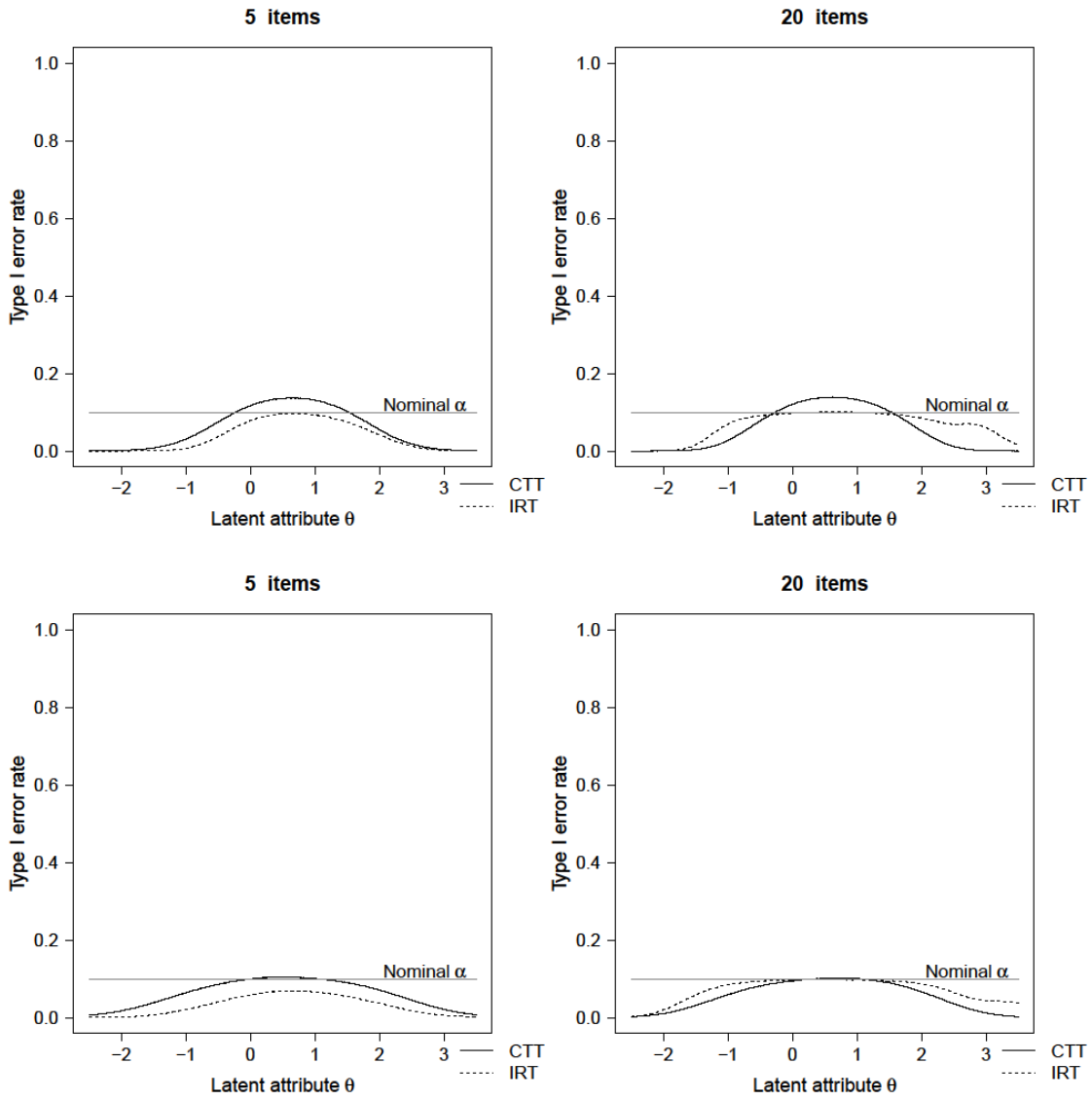


Figure 2. *Type I error rates in the homogeneous (upper panel) and heterogeneous (lower panel) item-difficulty conditions.*

Moreover, the asymptotically derived standard errors in IRT tend to overestimate the SE, particularly in the tails of the θ scale. This effect diminishes as the number of items increases. That is why increasing test length in IRT pushed the Type I error rates closer to the nominal Type I error rate across a wider range of the θ scale.

Detecting Improvement. For the population of truly improved persons, Table 2 shows the mean percentage of improved persons (i.e., $\delta < 0$, but change is not clinically significant) that both CTT and IRT detected. In general, differences between CTT and IRT were the largest for short tests ($J = 5$), large change ($\delta = -1.5$), and homogeneous item difficulties, where the highest mean difference was 18% in favor of CTT. Comparable detection rates were

Table 2. *Population-Level Classification Rates (Percentages Averaged Across 100 Replications) for Detecting Improvement in the Clinical Population, for Varying Item-Location Spread, Test Length and Test Model, and Three Cutoff Models.*

Cutoff	Homogeneous item difficulty						Heterogeneous item difficulty					
	5		10		20		5		10		20	
	CTT	IRT	CTT	IRT	CTT	IRT	CTT	IRT	CTT	IRT	CTT	IRT
Small change ($\delta = -.5$)												
a	13	8	24	22	40	41	7	4	13	13	22	26
b	18	13	28	25	44	43	11	6	15	15	24	27
c	8	6	18	18	34	37	4	3	9	10	18	23
Medium change ($\delta = -1.0$)												
a	29	20	54	51	75	78	17	10	33	35	56	64
b	41	32	63	60	80	79	23	15	37	37	58	64
c	18	11	40	39	64	69	9	6	23	24	45	55
Large change ($\delta = -1.5$)												
a	45	27	69	58	82	77	30	17	54	52	77	80
b	61	51	75	71	77	76	37	26	56	55	74	75
c	24	12	49	36	68	61	15	7.5	36	32	62	66

Note. Reliable change was tested at a nominal significance level of .10. Standard errors for differences between percentages ranged from 0.1% to 0.8%.

found for the three cutoff points. CTT had higher detection rates than IRT for 5 item-tests in all conditions and 10-item tests in the majority of the conditions; mean differences ranged from 1% to 18%. For the 20-item tests, IRT had higher detection rates than CTT in most conditions; mean differences ranged from 1% to 10%. For heterogeneous item-difficulty tests, on average CTT had higher detection rates for 5-item tests (mean difference ranged from 2% to 12%) and IRT for 20-item tests (mean difference ranged from 1% to 9%). Results were ambiguous for the 10-item condition. Increasing test length and the true change increased detection rates both for CTT and IRT.

Detecting Recovery. With respect to detecting recovery (i.e., clinical *and* statistical change) the largest differences between CTT and IRT with respect to mean detection rate were approximately equal to 12% (Table 3). For 5-item tests, CTT had higher detection of

Table 3. *Population-Level Classification Rates (Percentages Averaged Across 100 Replications) for Detecting Recovery in the Clinical Population, for Varying Item-Location Spread, Test Length and Test Model, and Three Cutoff Models.*

Cutoff	Homogeneous Tests						Heterogeneous Tests					
	5		10		20		5		10		20	
	CTT	IRT	CTT	IRT	CTT	IRT	CTT	IRT	CTT	IRT	CTT	IRT
Small change ($\delta = -.5$)												
a	22	17	33	29	49	47	13	9	18	18	28	31
b	6	3	12	15	23	35	8	4	12	14	19	26
c	23	18	34	31	49	48	14	10	19	20	28	31
Medium change ($\delta = -1.0$)												
a	47	42	68	67	82	82	27	21	41	44	63	68
b	16	9	34	38	59	70	16	11	29	33	47	59
c	48	40	68	68	81	83	29	22	44	47	63	69
Large change ($\delta = -1.5$)												
a	69	65	83	83	89	89	44	36	64	67	80	83
b	28	16	55	50	78	76	26	18	47	51	69	77
c	68	57	82	82	87	90	47	35	67	70	81	85

Note. Reliable change was tested at a nominal significance level of .10. Standard errors for differences between percentages ranged from 0% to 0.8%.

recovery than IRT across all levels of true change; differences varied between 2% and 13%. For 20-item tests, for the majority of the conditions IRT had higher detection rates than CTT; mean differences ranged from 2% and 13%. Results were consistent across homogeneous and heterogeneous item-difficulty tests and the three cutoff points (a, b, and c). For 10-item tests, results were ambiguous. In some conditions, CTT produced better detection rates than IRT and vice versa in other conditions. Again, increasing test length and true change increased detection rates for both CTT and IRT.

3.4 Discussion

A thorough methodological comparison of CTT and IRT with respect to individual-change assessment was absent thus far. We conclude that IRT is superior to CTT provided

that tests contain, say, at least 20 items, but in general the differences between the two methods are small. For shorter tests, results are ambiguous and using CTT seems to be a good choice. Instead of recommending the exclusive use of IRT for individual-change assessment (e.g., Prieler, 2007), we safely conclude that CTT and IRT each have their own advantages and disadvantages in different testing situations.

In order to minimize the burden on patients, shorter tests containing, say, approximately 5 items, may be preferred in clinical settings (e.g., Kruyen, et al., 2014). Here, CTT seems to better detect change than IRT, but one may notice that detection rates in the change conditions (i.e., $\delta < 0$) should be interpreted taking into account the empirical Type I error rates in the no change conditions (i.e., $\delta = 0$). For short tests, for homogeneous item-difficulty tests the (unknown) empirical Type I error rates generally were higher for CTT than for IRT, and in the middle of the θ scale Type I error rates were just above the nominal α level. Thus, for short tests IRT suggests individual change less frequently than CTT. This may partly explain why CTT more readily identifies improvement or recovery. On the other hand, since psychotherapies are meant to bring about positive change, the occurrence of zero true change in patients is rare in practice, thus causing Type II errors (i.e., concluding a patient did not change when in fact they did) to be more of a concern than Type I errors.

In general, because for short tests both in CTT and IRT the detection rates were generally low (below 50%) when true change was small ($\delta = -0.5$) or medium ($\delta = -1$), we do not recommend using short tests if such small changes are deemed clinically important. For large true change, detection rates were higher but a true change of this magnitude may be rare in practice. Future research may focus on empirical applications of IRT-based change assessment to gain more insight in the typical effect sizes. To summarize, we recommend using (1) tests containing at least 20 items and (2) IRT for scoring the tests. However, if the time and resources for administering longer tests are unavailable, we recommend using CTT which has more power when using short tests for detecting change in individuals. Another alternative based on IRT methodology is to use adaptive testing (Finkelman, Weiss, & Gyenam, 2010). In adaptive testing, the questionnaire is tailored to the current level of functioning whereby extreme scores due to floor or ceiling effects can be avoided.

An underexposed aspect of IRT-based individual-change assessment is its dependence on the fit of the model to the data. The greater the misfit between IRT model and data, the less accurate individual-change assessment is. Empirical evidence (Meijer & Baneke, 2004;

Sijtsma, Emons, Bouwmeester, Nykliček, & Roorda, 2008; Waller & Reise, 2010; Woods, 2006) suggests that traditional parametric models such as the GRM may be too restrictive to accurately describe clinical questionnaire data. Little is known about the robustness of individual-change assessment against model violations. CTT methods may be less sensitive to misfit and thus be a safer choice when IRT model-fit is questionable or inadequately demonstrated. Moreover, IRT methods require substantial samples sizes to obtain accurate parameter estimates, rendering CTT a justifiable alternative when samples are smaller. Second, most clinicians are familiar with basic CTT concepts such as reliability and SEM, but they lack sufficient knowledge of IRT. Because individual-change assessment also serves as a way to communicate between the clinician and the patient (Carlier & Roubertoux, 2010), it is important that the measurements used have a clear meaning to all parties involved.

The RCI, whether defined under CTT or IRT, assumes uncorrelated measurement errors within individuals. However, when errors are positively correlated, RCI values are too low. Such conservative RCI estimates are not necessarily problematic, because researchers maintain control over the Type I error rates and power which may remain sufficient to detect clinically relevant change. Furthermore, for low-stakes decisions (e.g., monitoring individuals throughout treatment), loss of power can be partly compensated by using a higher α level (e.g., $.10$ instead of $.05$). When measurement errors are negatively correlated, RCIs are overestimated. Such liberal RCIs are problematic, because they may overestimate treatment effects. Hypotheses about measurement errors across time can be tested using covariance structure analysis (e.g., Singer & Willet, 2003, p. 285). In case of anticipated negatively correlated errors, one should avoid using the RCI. More importantly, when measurement errors correlate positively across time, it is questionable whether individual-change scores can be interpreted meaningfully at all, because they may suggest lack of measurement invariance or undesirable idiosyncratic responding.

We used clinical cutoffs based on a fixed standardized difference of 0.5 between the functional and the dysfunctional populations. This resulted in cutoffs that were located either at the lower end, the middle, or the higher end of the latent-attribute scale. Consequently, the three cutoffs in our study represent clinical decisions at very different ranges of the latent-attribute scale rendering the results generalizable to many practical situations. In practice, populations may be further apart resulting in an increase of the standardized mean difference (Cohen's *d*), but then the JT cutoffs (*a*) (*b*) are pulled towards cutoff (*c*).

For detecting clinical change, we used the JT's approach which uses clinical cutoffs representing different levels of functioning. The use of clinical cutoffs for interpreting the clinical meaning of outcome measures is common. For example, clinical cutoffs are available for popular outcome measures such as the Hospital Anxiety and Depression Scale (Zigmond & Snaith, 1983), the Spielberger's State-Trait Anxiety Inventory (Spielberger et al., 1983), and the Qutcome Questionnaire-45 (Lambert et al, 1996; 2004). However, JT's method ignores the clinical relevance of change within either the clinical or functional ranges. Moreover, the cutoffs are based on sample data and thus are susceptible to sampling error. Therefore, another popular approach for operationalizing clinical significance is to define what constitutes minimum clinically important differences (MCIDs; e.g., Copay, Subach, Glassman, Polly, & Schuler, 2007; Wright, Hannon, Hegedus, & Kavchak, 2012). Observed change that exceeds this predetermined value is considered clinically relevant. A common choice for the MCID is a half standard deviation of pretest scores. However, research (Jabrayilov et al., 2015; Norman, Sloan, & Wyrwich, 2003) showed that change that is reliable also is clinically meaningful using an MCID based on the half standard deviation rule. Hence, results for detecting reliable change when true change is 0.5 in general also apply to detecting minimally clinical significant change of a half standard deviation.

To summarize, under ideal conditions (i.e., using simulated data) IRT produced better results than CTT with respect to individual-change detection, but differences generally were not as large as one would expect given the optimism the literature expresses (e.g., Prieler, 2007; Reise & Waller, 2009). Future research might compare CTT and IRT using real data, and might address issues like model fit and robustness of conditions when model fit fails.

Appendix

Computational Details for Deriving Clinical Cutoffs and Standard Error of Measurement

Let J be the number of items and $M + 1$ be the number of ordered answer categories. Furthermore, for the graded response model (GRM; Samejima, 1969) let $\boldsymbol{\xi}$ be the $J \times (M + 1)$ matrix of item parameters, for each item one slope parameter a and M thresholds b, and let *X* be the total score on the J items. Finally, we have *Q* latent attribute values θ_q ($q = 1, \ldots, Q$), where $Q = 500$, that are equidistant between -4 and 4. The conditional distribution of X given latent attribute value θ_q, denoted $f(X_+|\boldsymbol{\xi}, \theta_q)$, is a multinomial compound distribution (e.g., Kolen & Brennan, 1995). Because a closed form does not exist, the distribution is generated using a recursive algorithm (e.g., Emons et al., 2007; Thissen, Pommerich, Billeaud, & Williams, 1995). The *marginal* distribution of *X* is obtained as follows:

$$f(X|\boldsymbol{\xi}, \mu, \sigma^2) = \sum_{q=1}^{Q}\left[f\left(\mathrm{X}|\boldsymbol{\xi}, \theta_q\right) \times W_q(\mu, \sigma^2)\right], \quad \text{(A1)}$$

where $W_q(\cdot)$ are rectangular quadrature weights approximating the normal distribution with mean μ and variance σ^2 (e.g., Baker & Kim, 2004, p. 264). The mean of *X* in the population is defined by $E(X|\boldsymbol{\xi}, \mu, \sigma^2) = X \cdot f(X|\boldsymbol{\xi}, \mu, \sigma^2)$.

Clinical Cutoffs for Classical Test Theory (CTT)

For cutoff (*a*), we computed the marginal distribution of *X* (Equation $A1$) using quadrature weights from the standard normal distribution and as cutoff we selected the *X*-value closest to the 90th percentile; that is, $f(X \geq X_{\mathrm{cut}}|\boldsymbol{\xi}, \mu = 0, \sigma^2 = 1) \approx .90$. For cutoff (*b*), we computed the marginal distribution of *X* using quadrature weights from the normal distribution with mean 0.5 and variance of 1 and as cutoff we selected the X value closest to the 10th percentile of $f(X|\boldsymbol{\xi})$; that is, $f(X \leq X_{\mathrm{cut}}|\boldsymbol{\xi}, \mu = 0.5, \sigma^2 = 1) \approx .10$. For cutoff (c), we took the mean of the expected scores in both populations. Figure $1A$ of this supplement provides a graphical representation of the three cutoffs.

Standard Error of Measurement

Using quadrature weights based on the distribution in the clinical population, we computed the marginal item-score variances ($\sigma_{X_j}^2; j = 1, \ldots J$) and total-score variance (σ_X^2), and obtained Cronbach's alpha, which is defined as

$$\mathrm{alpha} = \frac{J}{J-1}\left[1 - \frac{\sum_{j=1}^{J}\sigma_{X_j}^2}{\sigma_X^2}\right].$$

The standard error of measurement (SEM) equals $\sigma_X\sqrt{1-\text{alpha}}$.

Computation of the Population-Level Type I Error Rates and Detection Rates

Let $S(\theta_q)$ be an outcome of interest conditional on θ_q. For example, $S(\theta_q)$ can be the detection rate for detecting a change of 0.5 for persons with θ_q at pretest. The population-level result for the clinical population is the weighted mean of the conditional results, $S(\theta_q)$, where the weights are the quadrature weights from the θ –distribution clinical population; that is,

$$S = \sum_{q=1}^{Q}\left[S(\theta_q) \times W_q(\mu = 0.5, \sigma^2 = 1)\right].$$

Population-level results for the healthy population are obtained in same way, but with quadrature weights from the θ-distribution in the healthy population.

Descriptive CTT Statistics for the Items and Tests used in the Simulation Study

Table A1: *Reliability, Item Means, and Item-Rest Score Correlations for the Items and Tests used in the Simulation Study.*

Descriptive Statistic	Clinical population					Healthy population				
	Mean	SD	Min	Max	IQR	Mean	SD	Min	Max	IQR
	Homogenous item difficulties / 5 items									
Cronbach's alpha	.83	.02	.78	.87	.02	.82	.02	.77	.86	.02
Item means	1.90	0.37	1.24	2.52	0.66	1.42	0.34	0.81	2.00	0.61
Item-rest correlations	.63	.04	.53	.71	.07	.62	.04	.52	.71	.07
	Homogenous item difficulties / 10 items									
Cronbach's alpha	.91	.01	.89	.92	.01	.90	.01	.88	.92	.01
Item means	1.88	0.36	1.24	2.52	0.62	1.40	0.33	0.81	2.00	0.58
Item-rest correlations	.66	.05	.57	.74	.08	.66	.05	.55	.74	.08
	Homogenous item difficulties / 20 items									
Cronbach's alpha	.95	.00	.95	.96	.00	0.95	.00	.94	.96	.00
Item means	1.86	0.37	1.24	2.52	0.64	1.38	0.34	0.81	2.00	0.59
Item-rest correlations	.69	.05	.59	.76	.08	.68	.05	.57	.76	.08
	Heterogeneous item difficulties / 5 items									
Cronbach's alpha	.68	.05	.58	.80	.07	.67	.05	.53	.80	.07
Item means	1.99	1.06	0.27	3.75	1.97	1.59	1.02	0.13	3.50	1.91
Item-rest correlations	.44	.08	.23	.63	.12	.43	.08	.22	.64	.13
	Heterogeneous item difficulties / 10 items									
Cronbach's alpha	.81	.02	.76	.86	.03	.80	.03	.73	.87	.04
Item means	2.00	1.07	0.27	3.74	2.02	1.60	1.03	0.13	3.49	1.90
Item-rest correlations	.49	.09	.29	.68	.14	.48	.09	.26	.69	.15
	Heterogeneous item difficulties / 20 items									
Cronbach's alpha	.90	.01	.88	.92	.02	.89	.01	.87	.92	.02
Item means	2.02	1.06	0.25	3.74	1.97	1.61	1.02	0.11	3.48	1.88
Item-rest correlations	.52	.09	.33	.71	.15	.52	.10	.31	.71	.17

Note. Results are based on 100 replications.

Graphical Display of Clinical Cutoffs in the Jacobson and Truax' (JT) Approach

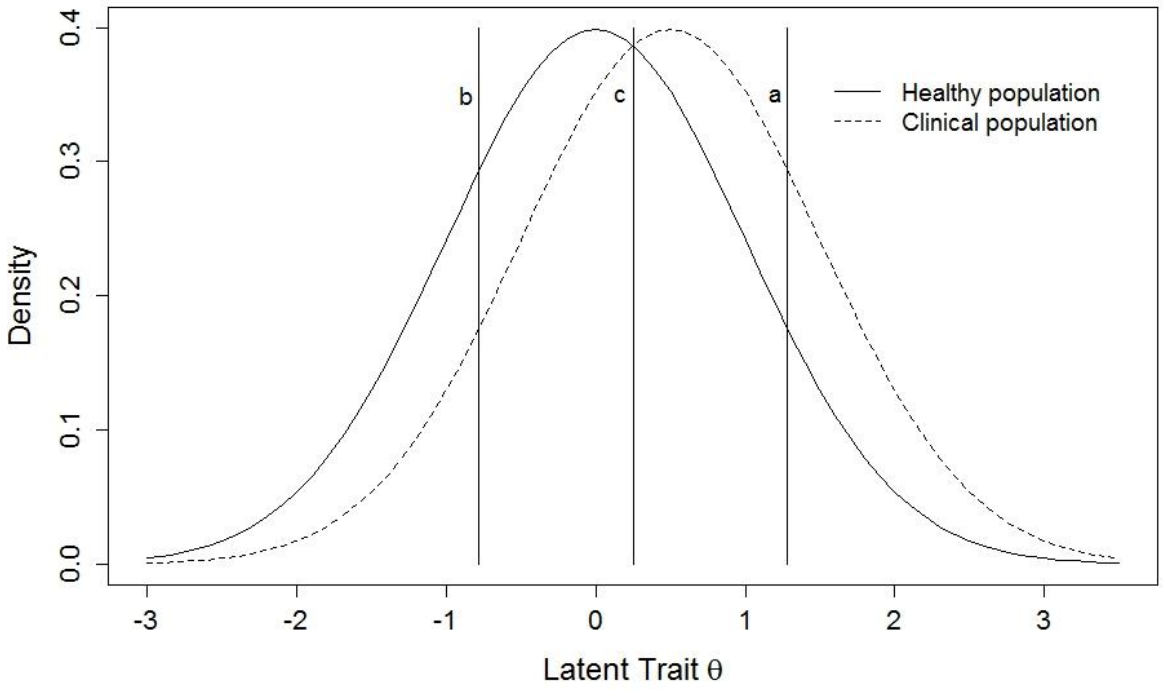


Figure A1. Cutoff points using hypothetical healthy and clinical population distributions. Distribution means equal 0 and 0.5, respectively, and the variances equal 1. Cutoff a is at the 90th percentile of the healthy population, cutoff b is at the 10th percentile of the clinical populations, and cutoff c is halfway the distribution means.

Chapter 4

Change Assessment Using IRT: An Illustration and Comparison with CTT-based Change Assessment

Abstract

Despite its popularity in the psychometric literature, in the practice of psychological testing item response theory (IRT) is rarely the preferred method of test-scoring and change assessment. Based on empirical data collected with the Outcome Questionnaire-45 in a Dutch outpatient sample, we demonstrate individual-change assessment using IRT and compare the results to the results obtained by means of classical test theory (CTT) methodology. In our study, compared to CTT, IRT was generally more likely to classify patients as having changed, that is, having improved or deteriorated. Moreover, results also show that test-scoring based on CTT and IRT is not the only factor affecting how and when patients are classified as having changed. Rather, test length and the method used for assessing the clinical importance of change also influence the conclusions about change.

Chapter 4

4.1 Introduction

Assessing change at the individual level is an essential part of clinical research and psychotherapy. Treatments having a modest mean effect at the population level may have a substantial positive or negative effect on individual patients (Hiller, Schindler, & Lambert, 2011; Jacobson & Truax, 1991). More importantly, assessment of patients on psychological outcomes such as anxiety, social well-being and general distress is becoming more and more popular in clinical practice. These so-called routine outcome measurements (ROMs) are assumed to provide clinicians and patients with important feedback on how they responded to the treatment. Research (e.g., Shimokawa, Lambert, & Smart, 2010; see also Boswell, Kraus, Miller, & Lambert, 2013) showed that such individual-level feedback increases treatment effectiveness and allows timely detection of deterioration.

We address two questions when assessing individual change, thus acknowledging that additional issues exist that we do not address here. An example is the validity of the measurement and possible shifts in the interpretation of test scores between two different measurement points. This is an important issue that we address elsewhere (Jabrayilov, Emons, & Sijtsma, 2015). The first question we address here is whether observed change reflects real change or whether it is also caused by measurement error? The answer to this question establishes whether observed change is *statistically* significant. Psychological outcome measurements are typically obtained by means of (self-report) questionnaires, which usually comprise only a limited number of items to minimize the burden on respondents (Boswell et al., 2013; Meier, 2008). Because these measurements are distorted by random measurement error, conclusions about patients derived from observed change scores are uncertain. Uncertainty can be substantial for many patients when short scales containing fewer than, say, ten items are used (Emons, Sijtsma, & Meijer, 2007; Kruyen, Emons, & Sijtsma, 2014). Therefore, clinicians should take the precision with which individual change has been measured into account when drawing conclusions about it. The second question is whether the observed change is clinically important or perceived as worthwhile by the patient. Low change scores may be statistically significant when reliably measured, but irrelevant for the person's functioning in daily life. To facilitate the assessment of clinical importance of change, clinicians need a frame of reference, such as cut-off scores that distinguish different levels of functioning.

Change Assessment with IRT: An Illustration and Comparison with CTT

So far, individual-change assessment has been predominantly based on psychometric methods from classical test theory (CTT; Lord & Novick, 1968). The best example from CTT is the widely-used reliable change index (RCI; Jacobson & Truax, 1991). Change assessment based on item response theory (IRT; Van der Linden & Hambleton, 1997; Reise & Revicki, 2015; Thomas, 2011) has been advocated as being superior to CTT approaches, because it uses finer-grained estimates of individual-level change resulting in higher precision (e.g., Prieler, 2007; Reise & Waller, 2009; Wise, 2004; Thomas, 2011). However, application of IRT to real data is complex and far from straightforward. Issues having received only little attention are parameter estimation, model-fit assessment, and assessment of clinical significance compared to statistical significance. The gap between the complexity of theoretical expositions of IRT (see also Brouwer, 2013; Sijtsma, Emons, Bouwmeester, & Nyklíček, Roorda, 2008), however useful, and practical applications may explain why IRT methods are still not the mainstream methodology despite their promising features. To bridge this gap, we believe that there is a need for comprehensive, illustrative empirical studies addressing IRT-based change assessment that focus on various data-analytic aspects.

From a practical point of view, given its simplicity one may also make a case for CTT-based methods and study the degree to which these methods underperform in change assessment compared to the more elaborate IRT methods. A useful question is whether one can profit from the subtleties of IRT in psychological testing which is often geared toward coarse classification of individuals into two or only a few categories, without the need for high precision between the cutoff scores. A simpler and albeit less precise statistical framework such as CTT may function quite well when coarse categorizations are required. Simulation studies comparing IRT and CTT in change assessment performance report ambiguous results (e.g., Brouwer, 2013; Jabrayilov, Emons, & Sijtsma, 2015; Sebille et al., 2010), suggesting that the advantages of using sophisticated IRT methods instead of the simpler CTT methods are not self-evident (Lance & Vandenberg, 2008, chap. 2). Moreover, there is a lack of empirical research providing head-to-head comparisons of the performance of the two approaches when assessing individual change.

In this study, we used IRT-methods to assess real persons' change on the Outcome Questionnaire-45 (OQ-45; Lambert et al., 1996). The OQ-45 is a widely-used ROM questionnaire, which has been translated into several languages. The OQ-45 covers three functional domains, which are symptom distress, interpersonal relations, and social role.

Several studies (e.g., Lambert et al., 1996; Vermeersch, Lambert, & Burlingame, 2000; Vermeersch, et al., 2004) have suggested that the OQ-45 is well suited for measuring change. For example, Vermeersch et al. (2004) found that, compared to controls, the OQ-45 detected significantly more improvement in people receiving psychotherapy. Using the OQ-45 and its subscales, we illustrated how IRT facilitates change assessment at the individual level and discuss important issues regarding practical implementation of IRT methods. In addition, using the same data we examined the extent to which CTT and IRT lead to different conclusions about individual change. We discuss the future use of CTT- and IRT-based individual-change assessment, and their application to individual change assessment with OQ-45.

4.2 Method

Participants

We conducted secondary data analysis using observations from $n = 540$ outpatients and $n = 1{,}807$ members of the general population (De Jong et al., 2007; Timman, De Jong, & De Neve-Enthoven, 2016). Clinical data from the outpatients came from three treatment departments within two medium-sized mental healthcare institutions in the Netherlands. These institutions treat members from an outpatient population with respect to a wide range of psychiatric disorders, including mood, anxiety, adjustment and personality disorders. The patients underwent therapy and completed the OQ-45 3.78 times on average (Minimum = 1 session, Maximum = 13 sessions, Mdn = 3 sessions). In total 81 patients completed the OQ-45 just once. Data from the general population were collected 1) in various businesses, 2) by contacting respondents through a phone book and 3) by private research institution TNS-NIPO. It constitutes a representative sample of the Dutch population with respect to sex, age, social economic status and education (De Jong et al., 2007). The sample from the general population is also referred to as the non-clinical sample.

Throughout this study, a distinction is made between the *calibration sample*, which is used to estimate the IRT models, and the *study sample*, which is used to empirically compare CTT and IRT approaches to change assessment. The calibration sample included data from the general population and the pretest data in the clinical sample (i.e., $n_{\text{calib}} = 540 + 1807 = 2347$). The study sample included the patients who completed OQ-45 at least twice

during therapy, resulting in a sample of $n = 540 - 81 = 459$ cases. In particular, we used as pretest data the OQ-45 scores at the very first session and as posttest data the OQ-45 scores at the very last session. Across the different subscales, 22, 28 and 25 patients had one or more missing scores in the SD, IR and SR subscales respectively. IRT approaches can handle missing data but CTT approaches to assessing individual change require complete data to obtain interpretable change scores because pretest and posttest total scores based on different number of items cannot be directly compared. Therefore, patients with missing observations were excluded from the analyses.

The Outcome Questionnaire-45 (OQ-45)

The OQ-45 (Lambert et al., 1996; see also De Jong et al., 2007) comprises 45 questions divided among three subscales, where each subscale covers a different functional domain. The *Symptom Distress* (SD) subscale consists of 25 questions that tap into symptoms of the most common types of psychological distress, in particular anxiety and depression. Examples of SD questions include "I fear fearful" and "I feel worthless". The *Interpersonal Relations* (IR) subscale consists of 11 items and measures to what extent respondents perceive difficulties in their relationships with family, friends and significant others. Example questions include "I am concerned about my family troubles" and "I have an unfulfilling sex life". Finally, the *Social Role* (SR) subscale consists of nine items that tap into dissatisfaction, distress, and conflict with one's employment, education and leisure activities. Example questions include "I feel stressed at work/school" and "I enjoy my spare time". Respondents answer the questions with respect to the past week using a 5-point rating scale ranging from 0 (*never*) to 4 (*almost always*). Positively worded items were reverse scored such that higher OQ-45 scores reflect worse functioning.

Two remarks are in order. First, we used data from the Dutch version of the OQ-45 (De Jong, Lambert, Nugter, & Burlingame, 2009). Psychometric structure, such as reliability and factorial structure, of the Dutch version were extensively studied (De Jong et al., 2007). Low loadings obtained from an exploratory factor analysis suggested that item 11 (i.e., "After heavy drinking, I need a drink the next morning to get going"), item 26 (i.e., "I feel annoyed by people who criticize my drinking or drug use"), and item 32 (i.e., "I have trouble at work/school because of my drinking or drug use") did not fit well into their subscales. In addition, low frequencies for scores 3 and 4 caused estimation problems for IRT models

(Conijn, Emons, De Jong, & Sijtsma, 2015). Hence, we excluded items 11, 26 and 32 from the analyses.

Second, the 25-item SD scale contains a 12-item subset known as the Anxiety and Somatic Distress (ASD) scale, reflecting that several SD items measure different dimensions of functioning. De Jong et al. (2007) suggested using the ASD scale as an additional clinically relevant subscale. However, the ASD scores explain 86% of the SD-score variance, suggesting high substantive overlap but lower reliability due to smaller test length. Because clinicians consider the ASD meaningful, we decided including the ASD scale in the study.

IRT Modeling of OQ-45

For the IRT analyses, we used the graded response model (GRM; Samejima, 1969). Let θ denote the latent variable of interest (e.g., social distress). The GRM assumes that θ is unidimensional and that variation in θ fully explains the association between the items; this property is technically known as the local independence assumption. Furthermore, let the OQ-45 items be indexed by *j*, such that $j = 1, \dots, 45$. The GRM describes each item by means of four cumulative logistic response functions, which are defined by a slope parameter (a_j) common to all functions, and four threshold parameters (b_{jm}) one for each function; that is, b_m $(m = 1, \dots, 4)$. The slope (a_j) expresses how well the item distinguishes low and high θ values. The threshold (b_{jm}) is the θ-value where the probability of scoring m or higher passes .50. For example, $b_{j2} = 1.2$ means that for item j persons with $\theta \geq 1.2$ are more likely to score at least 2 than below 2.

Ideally, the GRM parameters are estimated in a sample from the target population. However, for some items the extreme answer categories (i.e., 0 = *never*, or 4 = *almost always*) are rarely chosen. Consider an item on depressive thoughts for example. In a sample from the clinical population, respondents are unlikely to respond in the lowest categories which reflect near absence of depressive thoughts. As a consequence, the estimation of the threshold parameters for these extreme categories proves to be difficult because the parameters have to be estimated on the basis of highly sparse data (e.g., Hill et al., 2007). Estimating IRT parameters by collapsing answer categories to obtain a higher compound category score-frequency is technically possible but undesirable. This is due to the fact that using ad hoc collapsed item scales may reduce the scales' sensitivity to detect change at the extremes of the θ scales for future patients. Therefore, to avoid collapsing answer categories,

we constructed a calibration sample by combining the clinical sample and the non-clinical sample to obtain adequate item-score frequencies. Next, we used multiple-group IRT estimation (Bock & Zimowski, 1997) to estimate the GRM; multiple-group IRT corrects item parameter estimates for the heterogeneous composition of the sample.

Another important concern in IRT applications is model fit, especially when individual decisions are high-stakes. Conclusions derived from the estimated GRM can be trusted only when the IRT model fits the data. Trustworthiness of conclusions when the GRM does not fit is difficult to assess, because robustness results are often unknown; hence, we address GRM model-fit. Causes of misfit include (1) multidimensionality and/or local dependencies; and (2) misspecification of the functional shape between the latent variable θ and the cumulative item response probabilities. Comprehensive statistical tests for the GRM do not exist; instead, model fit has to be inferred from a combination of different goodness-of-fit tests, each focusing on different model assumptions. We evaluated model fit as follows. First, we compared the pair-wise observed inter-item correlations with those implied by the GRM. Discrepancies in excess of .15 (Morizot, Ainsworth, & Reise, 2007) indicate local dependencies. Second, we graphically inspected the observed response functions so as to check whether they exhibited the logistic S-shape typical of GRM cumulative response functions (Drasgow, Levine, Tsien, Williams, & Mead, 1995).

IRT and CTT Approaches to Individual-Level Change Assessment using QO-45

Testing change for statistical significance. Jacobson and Truax (1991; henceforth denoted JT) used CTT to formalize statistical and clinical significance of individual change. JT proposed the RCI to assess whether observed change is statistically significant. Let X_{pre} and X_{post} be the total score (i.e., the sum of the item scores) at pretest and posttest, respectively; and let $d = X_{\text{post}} - X_{\text{pre}}$ be the change score. Because higher OQ-45 scores reflect worse functioning, negative change scores ($d < 0$) reflect improvement. The RCI is defined as

$$RCI = \frac{d}{SEM_d}, \tag{1}$$

where SEM_d is the standard error of measurement of the difference score d. Statistic SEM_d describes random variation of the change scores d when true change is absent. In practice, SEM_d is often computed as $\sqrt{2}SEM_{\text{pre}}$ (Jacobson & Truax, 1991), where SEM_{pre} is the standard error of measurent for pretest scores. The RCI is assumed to be standard normally distributed under the null hypothesis of no change. Using a critical value of the *Z*-distribution,

one may test whether change is significant. For example, $|RCI|$ > 1.96 means change is significant at the 5% level (two-tailed). Significant change is also known as *reliable change*.

In practice, the RCI is often defined differently as the minimum number of raw score points to be gained or lost to be qualified as a reliable change. For example, the RCIs for the Dutch OQ-45 are defined as 55 score points on the total scale, and 14, 10, 8 and 9 score points for the SD, IR, SR and ASD subscales, respectively (De Jong et al., 2007). These RCI criteria were based on the complete set of 45 items; however, due to the fact that we discarded three items these proposed RCIs were not applicable in our study. Therefore, we computed the RCIs (Equation 1) using the sample estimates of SEM_d for the scales from which the poor fitting items were removed.

Generalization of the RCI to an IRT context is straightforward (e.g., Reise & Haviland, 2005). In IRT, persons are measured on the latent variable θ-scale by means of the estimated $\hat{\theta}_{\mathrm{pre}}$ and $\hat{\theta}_{\mathrm{post}}$ scores, so that the estimated change score equals $\hat{\delta} = \hat{\theta}_{\mathrm{post}} - \hat{\theta}_{\mathrm{pre}}$. The standard error of $\hat{\delta}$ is obtained using Fisher information (e.g., Embretson & Reise, 2000; Kruyen et al. 2014). Let $I\left(\hat{\theta}_{\mathrm{pre}}\right)$ and $I\left(\hat{\theta}_{\mathrm{post}}\right)$ be the information at $\hat{\theta}_{\mathrm{pre}}$ and $\hat{\theta}_{\mathrm{post}}$. The standard error equals

$$SE(\hat{\delta}) = \sqrt{\frac{I(\hat{\theta}_{\mathrm{pre}}) + I(\hat{\theta}_{\mathrm{post}})}{I(\hat{\theta}_{\mathrm{pre}})I(\hat{\theta}_{\mathrm{post}})}}.$$

(see Kruyen et al., 2014), and the IRT-version of the RCI equals

$$RCI_{IRT} = \frac{\hat{\delta}}{SE(\hat{\delta})}.$$

The RCI_{IRT} is assumed to be standard normally distributed given absence of change.

The standard error of the change, $SE(\hat{\delta})$, reflects the precision with which change is assessed and depends on the location on the θ-scale where the change has occurred. Consequently, persons located at different θ values and showing the same absolute observed change $\hat{\delta}$ may produce different RCIs, which can lead to different conclusions about reliable change. For example, in IRT a change from 0.5 to 0.2 $(\hat{\delta} = -0.3)$ may be significant, but a change from 1.5 to 1.2 (i.e., $\hat{\delta} = -0.3$) may not. In contrast, CTT uses the same standard error of measurement for all persons, so that, without exception, all equal change scores are either significant or insignificant. As a result, several authors (Embretson and Reise, 2000; Prieler, 2007; Reise & Haviland, 2005) argued that CTT is unrealistic and that IRT methods should be preferred.

Clinical significance: JT approach. To assess whether change is clinically significant, JT proposed dividing the total-score scale into a functional range and a dysfunctional range. Patients showing reliable change, whose OQ-45 score moves from the dysfunctional into the functional range, are considered to be *recovered*. When patients showing reliable change, move from the functional into the dysfunctional range, JT speak of *clinical deterioration*. Patients showing reliable change but who not pass the cutoff value, are said to have *improved* when change is positive, and *deteriorated* when change is negative. JT proposed three methods to determine clinical cutoff values for distinguishing functional from dysfunctional ranges, all based on a normative sample from a functional population, a dysfunctional population, or both. A detailed explanation of the JT methods and alternative methods is beyond the scope of this paper.

JT-based clinical cutoff values also exist for the Dutch OQ-45. In particular, clinical cutoffs (denotes as X_{cut}) equal 55, 33, 12, 10, and 19 for the total scale, and the SD, IR, SR, and ASD subscales, respectively (De Jong et al., 2007). The original clinical cutoff values were converted to clinical cutoff values on the shortened scales. The new cutoff value is the total score on the shortened test that has the same percentile rank as the original cutoff value for the full-length test. The newly derived cutoff values were 33, 12, 9, and 19, for the subscales SD, IR, SR and ASD, respectively.

We implemented JT's approach to clinical significance in the context of IRT, by finding a cutoff value on the θ scale (θ_{cut}) which distinguishes the dysfunctional and functional ranges, comparable to the CTT value. JT's methods for deriving clinical cutoff values can also be applied to IRT scores $\hat{\theta}$, but to ensure that cutoffs θ_{cut} correspond to the Dutch clinical OQ-45 cutoff values for total scores (denoted by X_{cut}) we obtained the θ_{cut} values as follows. Let $E(X|\theta, \boldsymbol{\xi})$ be the expected total score under the postulated IRT model with item parameters $\boldsymbol{\xi}$ and person parameter θ. For θ_{cut}, we chose the value of θ for which the equality $E(X|\theta, \boldsymbol{\xi}) = X_{\text{cut}}$ holds. Figure 1 illustrates the transformation of the cut scores on the X scale (y-axis) into cut scores on the θ scale (x-axis) for the Social Distress scale (detailed discussion follows). In this example, the clinical cutoff value of 33 corresponds to cutoff value -0.99 on the θ-scale.

Clinical significance: Minimum clinically important difference. Sometimes clinicians also use rules of thumbs for the minimum change that is considered clinically significant; this is the minimal clinically important difference (MCID; e.g., Copay et al., 2007). A frequently used

MCID is the "half standard deviation rule" (Norman, Sloan, & Wyrwich, 2003). This rule defines change to be clinically significant if the change exceeds half the standard deviation of the pretest scores in the clinical population. The "half standard deviation rule" seems to be in a widespread use, thus ignoring the specific clinical context envisaged, where different contexts could require different MCIDs (Wright, Hannon, Hegedus, & Kavchak, 2012). For example, using MCIDs based on this rule can underestimate the practical importance of change in contexts where even small change on the attribute has a meaningful impact on a patient's daily functioning, or it can overestimate the clinical significance of change when only larger differences are meaningful. Therefore, we compared CTT to IRT for three levels of the MCID to study the sensitivity of the results to different MCIDs. We studied the following three rules: Let $S_{X_{\text{pre}}}$ be the standard deviation of the pretest scores in the clinical population, then MCID = $0.2 \times S_{X_{\text{pre}}}$ (small effect), MCID = $0.5 \times S_{X_{\text{pre}}}$ (medium effect), and $\text{MCID} = 1 \times S_{X_{\text{pre}}}$ (large effect).

Statistical Analysis and Software Used

We compared IRT and CTT with respect to classification of individual patients into different categories of individual change. A distinction was made between classifications following the JT approach and those obtained using the MCID criteria. For the JT approach, each patient was classified into one of the following five exhaustive and mutually exclusive categories: (i) Reliable and clinically significant deterioration (RC Det); (ii) Reliable deterioration (R Det); (iii) No reliable change (NC); (iv) Reliable improvement (R Imp); and (v) Reliable and clinically significant improvement (RC Imp). Following Emons et al. (2007), who argued that for high-stakes decisions certainty levels of .90 or higher are deemed acceptable, we used a 10% significance level (two-tailed) for testing the null hypotheses of no reliable change. For clinical change using MCID criteria we classified patients as having changed (i.e., change score exceeds the MCID) or clinically unchanged (i.e., change score is less than the MCID), irrespective of whether change was reliable.

GRM item parameters were estimated using multiple group MML estimation in FlexMIRT (Edwards & Cai, 2013; syntax available upon request). The θ distribution in the clinical population was restricted to be standard normal so as to statistically identify the latent variable scale. The mean and variance of the θ distribution were estimated simultaneously with the item parameters. Individual person parameters were estimated

using weighted maximum likelihood (WML; Warm, 1989). Because WML estimates are not available in FlexMIRT, we developed our own software in C++ (available upon request from the second author), which utilizes Newton-Raphson optimization.

4.3 Results

Sample Statistics and Mean Treatment Outcomes

For each of the subscales, Table 1 shows the descriptive statistics and CTT statistics in the non-clinical and the clinical samples (pretest and posttest data).

Table 1. *Summary Statistics of Subscale Total Scores in the General Population and the Clinical Sample.*

	Population		
	Non-Clinical	Clinical Pretest	Clinical Posttest
	Symptom Distress (SD) Subscale (24 items)		
n	1767	437	437
mean X_+	23.63	47.63	41.11
SD X_+	11.45	14.98	16.49
Alpha	.90	.91	.94
SEM	3.54	4.37	4.02
	Interpersonal Relations (IR) Subscale (10 items)		
n	1773	431	431
Mean	8.83	15.44	14.43
Std.	4.87	6.06	6.58
Alpha	.72	.76	.83
SEM	2.57	2.98	2.71
	Social Relations (SR) Subscale (7 items)		
n	1779	434	434
Mean	7.13	10.70	9.62
Std.	3.57	4.84	4.66
Alpha	.63	.70	.73
SEM	2.16	2.64	2.42
	Anxiety and Somatic Distress (ASD) Subscale (12 items)		
n	1786	444	444
Mean	12.82	23.56	20.34
Std.	6.43	8.20	8.87
Alpha	.83	.84	.89
SEM	2.64	3.26	3.00

Note. Std. = standard deviation; Alpha = coefficient alpha; SEM = standard error of measurement. Item 11 (SD), item 26 (IR), items 12 and 32 (SR) were excluded from the OQ-45 due to low scalability within the subscale.

Except for the SR scale in the non-clinical sample, for all subscales coefficient alpha exceeded .70. Interestingly, even though alphas were lower in the non-clinical samples, most standard errors of measurement (SEMs) were smaller in the non-clinical sample. Smaller SEMs may be caused by restriction of range of OQ-45-scores in the non-clinical sample due to floor effects, which reduces the error variance and thus the SEM (Lord & Novick, 1968, p. 233). Results further show that reliability was higher at posttest than at pretest. This may be have been caused by increased sample heterogeneity at posttest (Lord & Novick, 1968, pp. 198-199); i.e., some patients may have changed more than others resulting in more variation in OQ-45 scores after the treatment than before it.

For each subscale, standardized mean differences between the non-clinical and pretest clinical samples were significant at the 1% significance level (independent samples *t*-test, two-tailed) and large (Cohen's d values > 1). Mean differences showed that the clinical conditions of patients tended to improve between pretest and posttest. They were significant at the 1% significance level (paired samples *t*-test, one-tailed test). Standardized treatment outcomes were estimated to be small for IR $(d = -0.16)$ and SR ($d = -0.16$), and moderate for SD ($d = -0.41$) and ASD ($d = -0.38$). In the absence of control-group information, conclusions about causality with respect to the effect of the treatment may be imprudent. For example, individual improvement may also reflect a statistical artifact known as regression to the mean. For this design, it is impossible to distinguish real effects from artifacts for individuals.

IRT Modeling of the OQ-45

SD scale. For each subscale, Table 2 shows mean estimated item parameters and their ranges. Figure 1a shows the expected total-score and the test-information functions. The vertical dashed line in Figure 1a denotes the clinical cutoff; respondents whose $\hat{\theta}$ value reliably moved from the dysfunctional into the functional range clinically improved. Results show that the SD scale is most informative at clinical ranges of the θ scales. Residual inter-item correlations ranged from $-.08$ to $.09$, which were below the critical threshold of .15. For the SD scale, we additionally employed confirmatory bifactor modeling (Reise, Morizot, & Hays, 2005) to evaluate whether unidimensional IRT modeling is justified even though a subset of items may

Table 2. *Mean α and b Parameter Estimates for the SD, IR, SR and ASD Subscales.*

OQ-Scale	α	b_1	b_2	b_3	b_4
SD	1.62	-2.19	-0.66	0.77	2.80
	(0.67; 2.82)	(-4.18; -0.01)	(-1.91; 1.18)	(-0.45; 2.39)	(1.15; 4.14)
IR	1.42	-1.85	0.08	1.80	3.36
	(0.55; 2.57)	(-4.54; -0.48)	(-1.61; 1.53)	(0.73; 3.36)	(2.03; 5.29)
SR	1.71	-1.09	0.35	1.49	2.65
	(0.90; 3.04)	(-1.71; 0.00)	(-0.76; 1.19)	(0.47; 2.25)	(1.55; 3.73)
ASD	1.68	-1.47	-0.18	1.03	2.38
	(0.91; 2.43)	(-3.07; -0.02)	(-0.93; 0.91)	(0.28; 1.85)	(1.60; 3.33)

Note. SD: Symptom Distress; IR: Interpersonal Relations; SR: Social Role; ASD: Anxiety and Somatic Distress. Minimum and maximum parameter values are shown in parentheses. Latent variable θ is assumed to be normally distributed in the clinical population.

also load on a second ASD dimension (De Jong et al., 2007). Given the bifactor model, all SD items loaded on a general factor, whereas the subset of ASD items also loaded on a second factor that is independent of the general factor (i.e., the factors are orthogonal). The GRM item slopes were 0.86 to 1.17 times higher than the corresponding slopes for the general factor in the bifactor model. Hence, local dependencies had little effect on the estimated GRM slopes which, following Reise at al. (2007), justifies the use of the unidimensional GRM for the SD scale. Graphical item-fit analysis showed misfit with respect to the GRM logistic shape, but only for extreme θ values. We conclude that the GRM did not fit perfectly but well enough to consider its fit to the SD scale data adequate for the current application.

IR scale. Graphs of the expected total-score and the information functions (Figure 1) showed that the IR-scale is most informative in the clinical range. The residual correlation between the items 7 and 17 equaled .23, suggesting that local dependence was problematic. To evaluate the practical importance of local dependence, we followed De Cock et al. (2011; also, see Edwards & Cai, 2011) and estimated the slope of item 7 while excluding item 17, and vice versa. Estimated slopes decreased by 0.05 (item 7) and 0.04 (item 17), a difference small enough to suggest that local dependence had little negative effect on parameter estimates. The other inter-item residuals ranged between $-.16$ and

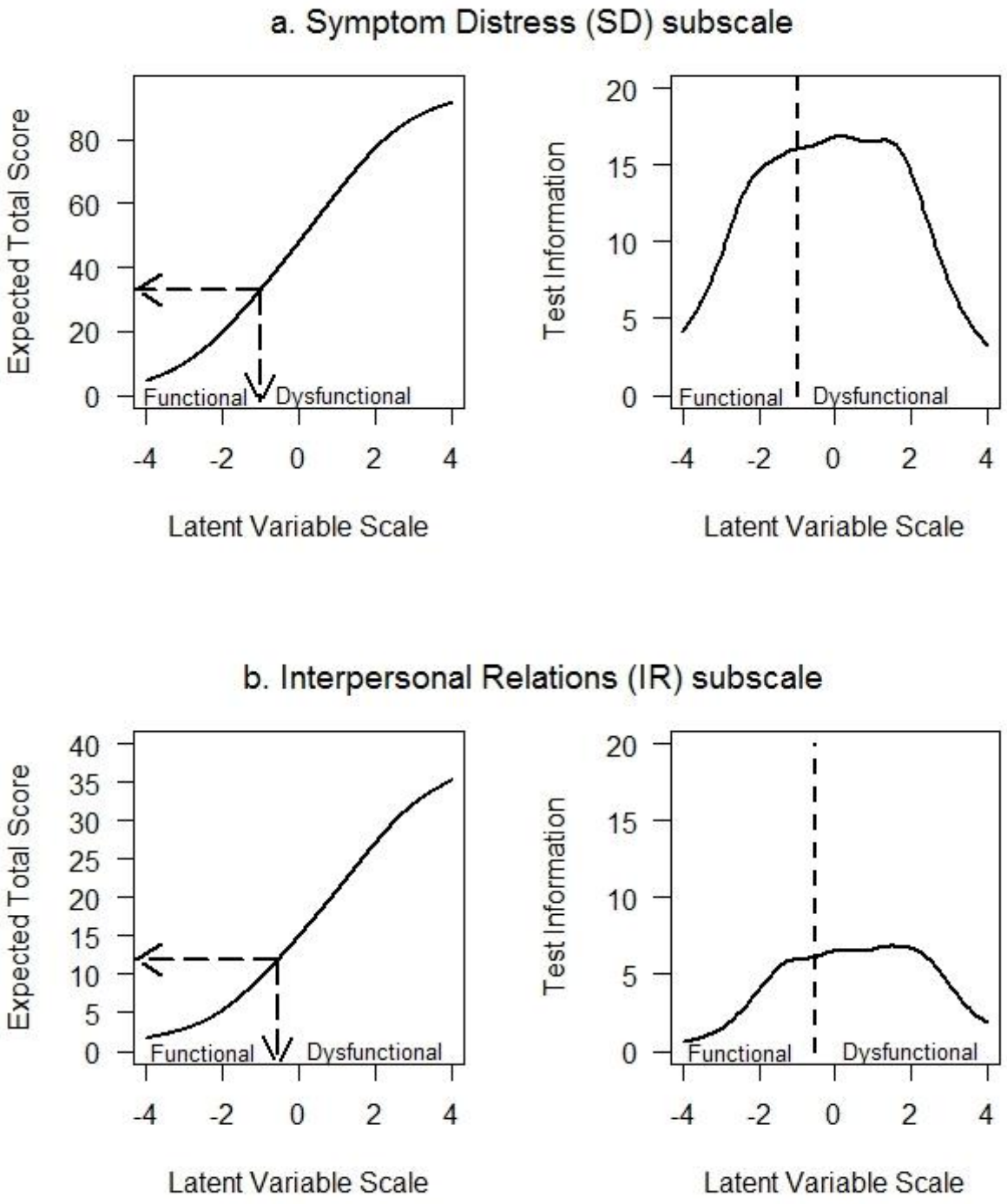
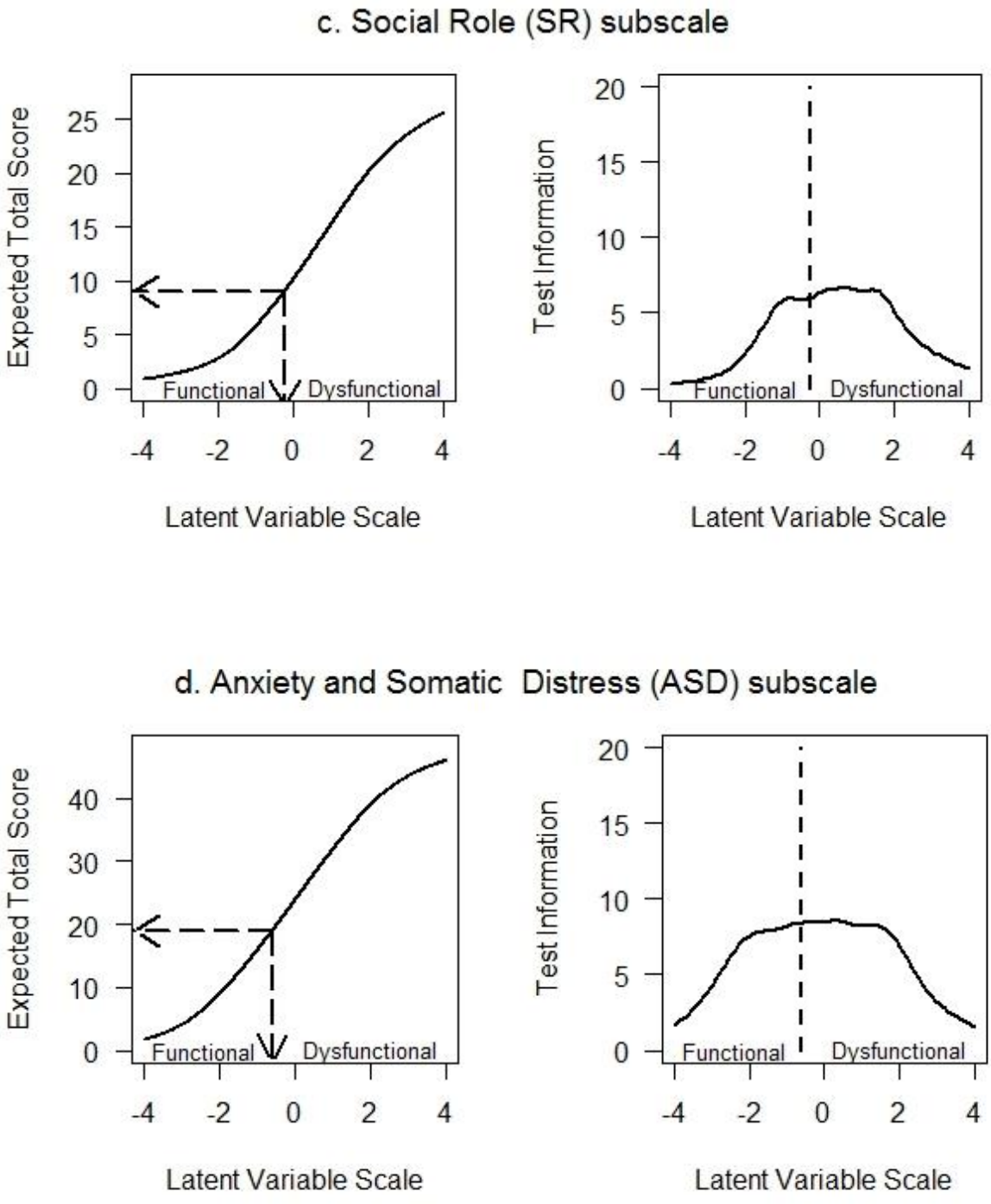


Figure 1. *Expected total-score functions and test-information functions. Dashed lines indicate Dutch clinical cutoffs.*

Figure 1 continued



.10. Graphical item-fit analysis showed that GRM expected item-response functions differed little from the observed curves nearly everywhere on the θ-scale, except for high θ values. We concluded that the items showed satisfactory fit for the current application.

SR scale. The fit of the GRM to the SR scale was questionable. First, graphical fit analysis suggested that item 12 had the largest misfit. Removal of this item produced better model fit but model fit was still questionable for the remaining items. Removal of additional items did not further improve the fit. In general, we conclude that SR may not be a psychometrically sound standalone test.

ASD scale. We recalibrated the item parameters for the ASD scale (Table 2). The ASD scale is most informative in the clinical range, but information was less than that of the SD scale. Because the GRM fitted well to the ASD items when these items were included in the SD scale, we concluded that the GRM fitted satisfactorily to the ASD items.

Comparing CTT and IRT Approaches

JT-approach to individual change assessment. Table 3 shows the cross tabulation of patient classifications for CTT and IRT based on the JT criteria of statistical and clinical significance. Diagonal boldfaced frequencies indicate classification agreement between CTT and IRT. For all four subscales, compared to the CTT approach, the IRT approach had more power to detect reliable change in patients. For example, for the SD subscale IRT classified 208 (i.e., 437- 229; 47.5%) and CTT classified 162 (i.e., 437-275; 37%) patients as having changed reliably in either direction. Inspection of the off-diagonal elements in Table 3 showed that IRT and CTT disagreed in decisions about reliable and/or clinical significance of change for 15% of the patients in the SD subscale, 11% in the IR subscale, 17% in the SR subscale, and 16% in the ASD subscales. Hence, CTT and IRT agreed in 85% of cases for the SD, 89% for the IR, 83% for the SR and 84% for the ASD subscales.

For each scale, IRT and CTT suggested drastically different classifications for a few patients, such as recovery (RC Imp or RC Det) versus no change (NC), but upon closer inspection we found that these patients showed change scores equal to 9 or 10, yielding CTT-based RCIs in the range $[-1.46; -1.62]$, which are just below the critical value of the RCI, and IRT-based RCIs in the range $[-2.25; -1.77]$, which are just above the RCI critical value. These results suggest that IRT more often detected subtle changes. Similar results were found for the other scales. Greater suitability of IRT for subtleties is what we expected.

Table 3. *Cross Classification of Patients Based on Jacobson and Truax's Criteria for Statistical and Clinical Significance of Change.*

	IRT-based					
	RC Det	R Det	NC	R Imp	RC Imp	Total
Symptom Distress (SD) Scale						
CTT-based:						
RC Det	**5**	0	0	0	0	5
R Det	2	**16**	2	0	0	20
NC	3	5	**227**	35	5	275
R Imp	0	0	0	**69**	8	77
RC Imp	0	0	0	6	**54**	60
Total	10	21	229	110	67	437
Interpersonal Relations (IR) Scale						
CTT-based:						
RC Det	**8**	0	0	0	0	8
R Det	1	**11**	3	0	0	15
NC	4	8	**326**	9	9	356
R Imp	0	0	6	**17**	2	25
RC Imp	0	0	2	5	**20**	27
Total	13	19	337	31	31	431
Social Role (SR) Scale						
CTT-based:						
RC Det	**3**	0	1	0	0	4
R Det	4	**4**	0	0	0	8
NC	13	5	**337**	17	15	387
R Imp	0	0	3	**16**	5	24
RC Imp	0	0	0	0	**16**	16
Total	20	9	341	33	36	439
Anxiety and Somatic Distress (ASD) Scale						
CTT-based:						
RC Det	**6**	1	0	0	0	7
R Det	2	**8**	2	0	0	12
NC	3	5	**278**	30	12	328
R Imp	0	0	1	**29**	6	36
RC Imp	0	0	0	2	**59**	61
Total	11	14	281	61	77	444

Note. RC Det = Reliable and clinically significant deterioration; R Det = Reliable but not clinically significant deterioration; NC= No reliable change; R Imp = Reliable but not clinically significant improvement; and RC Imp = Reliable and clinically significant improvement. Classifications in the same categories by CTT and IRT (diagonals) are shown in boldface.

MCID approach to change assessment. Table 4 shows the detection rates for clinically significant change using the MCID.

Table 4. *Detection of Clinically Significant Change Using Three Different Operationalizations of the Minimal Clinically Important Difference (MCID). Table Entries are Percentages.*

OQ-Scale	Detection Rate		Cross-Classification Detection MICD			
	CTT	IRT	CTT−/IRT−	CTT+/IRT−	CTT−/IRT+	CTT+/IRT+
			Small MCID			
SD scale:	85.1	77.8	13.3	8.9	8.9	76.2
IR scale:	69.6	75.6	16.2	8.1	14.2	61.5
SR scale:	89.2	76.6	7.6	15.7	3.2	73.5
ASD scale:	79.7	78.2	15.1	6.8	5.2	73.0
			Medium MCID			
SD scale:	52.9	50.8	43.5	5.7	3.7	47.1
IR scale:	40.4	43.6	47.1	9.3	12.5	31.1
SR scale:	50.5	49.5	39.2	11.3	10.4	39.2
ASD scale:	49.8	56.1	41.2	2.7	9.0	47.1
			Large MCID			
SD scale:	24.3	227	73.5	3.9	2.3	20.4
IR scale:	17.4	16.7	77.5	5.8	5.1	11.6
SR scale:	24.2	22.8	69.9	7.6	6.2	16.6
ASD scale:	20.9	25.7	72.7	1.6	6.3	19.4

Note. CTT− = No clinical change using CTT; IRT− = No clinical change using IRT; CTT+ = clinically significant change using CTT; IRT+ = clinically significant change using IRT. SD: Symptom Distress; IR: Interpersonal Relations; SR: Social Role; ASD: Anxiety and Somatic Distress.

In general, detection rates decreased as effect sizes increased. CTT showed higher detection rates when change was small (upper panel) and ambiguous results when change was medium or large. For the latter two MCID levels, absolute differences ranged between 1% (SR subscale) and 6.3% (ASD subscale), with CTT most often performing better than IRT. For the ASD scale, IRT performed better.

4.4 Discussion

In general, compared to CTT, for the OQ-45 scales the GRM was more effective in detecting reliable change, but results were ambiguous for detecting MCIDs. Because in a real-

settings the true status with respect to change is unknown and because psychometric models such as CTT and the GRM rest on assumptions that are approximations of the data structure at best, conclusions should be considered with caution. Based on the characteristics of the OQ-45 data, the authors did some simulations to investigate the extent to which results were replicable under known and ideal circumstances (details can be found in the Appendix). Results of the simulations supported the main findings with respect to differences between IRT and CTT to detect reliable change, but as one would expect absolute detection rates were higher in the simulations than in the empirical-data analysis.

Jabrayilov, Emons, and Sijtsma (2015) did an elaborate simulation study, and found smaller differences between IRT and CTT results with respect to change assessment compared to the differences found in this study. Differences depend on item and population characteristics. The item parameters of OQ-45 are well-spread across the θ scale. For large item spread, Jabrayilov et al. (2015) found that IRT performs better than CTT when tests contain at least 20 items. Results of the present study may have also been affected by the various choices we made. For example, following standard practice we computed the SEM_d using the estimated SEM from the pretest. Alternatively, SEM_d can be computed as a pooled SEM from pretest and posttest; that is, as $\sqrt{SEM_{\text{pre}}^2 + SEM_{\text{post}}^2}$ (e.g., Maassen, 2009). Like SEM_d, pooled SEM assumes equal SEMs at pretest and posttest, but it is less prone to sampling error (cf. the pooled variance used in the independent samples t-test). The results for both SEMs showed minor differences.

We used the unidimensional GRM, but many outcome questionnaires are inherently multidimensional because they address different facets of functioning (Reise & Revicki, 2015). Hence, more complex multidimensional IRT models (Reckase, 2009) or bifactor models (e.g., Reise, Morizot, & Hays et al., 2005) may be used. The use of more complex models comes at a price, however, because defining change in a multidimensional context is less straightforward due to the multitude of combinations of directions in which change might happen. In addition, more-complex models require larger samples for obtaining accurate parameter estimates. Finally, several authors have noticed that multidimensional IRT models involve conceptual peculiarities thus questioning their usefulness (Hooker, Finkelman, & Schwartzman, 2009; see Van der Linden, 2012, and Van Rijn & Rijmen, 2012, for a critical discussion).

Two suggestions for future research are the following. First, lack of measurement invariance (e.g., Millsap, 2010; Golembiewski, Billingsley, & Yeager, 1976) is a serious topic for change assessment. For instance, after treatment patients may perceive items differently than before, and similar responses before and after treatment may have a different meaning. Different meanings violate measurement invariance across time and invalidates change based on pretest and posttest scores. Therefore, future studies should investigate measurement invariance in the OQ-45. Second, because no IRT model fits data perfectly, there is always some degree of misfit and researchers often do not know what to do when it is present in large samples (e.g., Sijtsma, 2012). Researchers would benefit from a clearer set of guidelines to evaluate the size of misfit and possible solutions given the desired application.

To conclude, both CTT and IRT methods have their benefits. CTT is well-known, providing easy to use and robust guidelines for detecting individual change. Its simplicity helps the facilitation of communication between patient and clinician. IRT is more precise, and more flexible allowing technical extensions but less straightforward than CTT to apply to real data.

Appendix

For each patient in the clinical sample, we generated pretest and posttest data using his/her estimated pretest and posttest θ values. Item-score vectors were generated under the GRM, where each OQ-45 item was modeled by one slope parameter and four threshold parameters. In particular, let $\hat{\xi}_{jm}$ $(j = 1, \dots J, m = 1, \dots, 5)$ denote parameter estimates of item j, such that $\hat{\xi}_{j1}$is the estimated slope of item j, and $\hat{\xi}_{j2}, \dots, \hat{\xi}_{j5}$ are the estimated threshold parameters (summary statistics for the estimated parameters can be found in Table 2). Furthermore, let the $SE(\hat{\xi}_{jm})$s be the standard errors of the item parameters, which were obtained by means of FlexMIRT (not tabulated). Pretest data and posttest data were generated using item parameters that were randomly drawn from their posterior distribution, which for each parameter of item j is the normal distribution with mean $\hat{\xi}_{jm}$ and variance $SE^2(\hat{\xi}_{jm})$. Person parameters were re-estimated from the simulated item-score vectors using $\hat{\xi}_{jm}$. Hence, the item parameters used for simulating the data differed a little from the item parameters used to re-estimate the person parameters, where differences were defined by the standard errors. This way of simulating the data mimics the idea that in reality sample estimates of the items which contain estimation errors are used to estimate person parameters. Change was assessed by means of the RCI, using the SEMs from the empirical data analysis (Table 1) for CTT and the standard errors of the re-estimated person parameters for IRT. Data were simulated 100 times yielding 100 replications, each replication based on newly sampled item parameters. Table A1 reports the mean values across the replications.

Table A1: *Detection Rates of Reliable Change in the Simulation Study*

Scale	CTT		IRT	
	Empirical	Simulated	Empirical	Simulated
SD	37.1%	42.8%	47.6%	51.7%
IR	17.4%	25.3%	21.8%	31.1%
SR	11.8%	21.2%	22.3%	32.7%
ASD	26.1%	36.4%	36.7%	43.3%

Note. Mean values are based on 100 replications

Chapter 5

Examining Measurement Invariance in the Dutch Outcome Questionnaire-45

Abstract

In the absence of measurement invariance across time, change scores based on pretest-posttest measurements may be inaccurate representations of real change on the latent variable. Based on a combination of factor analysis and item response theory methodology, we examined measurement invariance in the Dutch version of Outcome Questionnaire-45 (OQ-45). Using secondary data analysis of a sample of $N = 540$ Dutch outpatients, we tested the stability of the factorial structure and the metrical invariance across pretest and posttest measurements. Results revealed stable factorial structure from pretest to posttest and minor violations of metric invariance for two items in the Dutch OQ-45. However, the effects of these violations on practical change assessment were negligible.

Chapter 5

5.1 Introduction

Assessing psychotherapy outcomes typically involves taking into account the difference between pretherapy and posttherapy scores on a self-report questionnaire, thus assuming that the test has invariant measurement properties across time. Violation of the assumption of temporal measurement invariance renders the meaning of change scores ambiguous, because it is no longer clear whether observed change is due to real change on the latent variable, or whether it is caused by other, irrelevant factors (Millsap, 2010; Schmitt, 1982). Research has also shown that questionnaires failing to demonstrate measurement invariance over time tend to have a poor reliability and predictive validity (e.g., Alvares & Hulin, 1972; Henry & Hulin, 1987).

The assumption of temporal measurement invariance is violated when the relationship between the responses and the latent variable changes over time. According to Golembiewski, Billingsley, and Yeager (1976) this relationship can change in two ways. The first type of change occurs when the respondents recalibrate the item response options at posttest. For example, at posttest a patient may perceive the response option “often being unhappy" to represent levels of unhappiness that are different than levels perceived at pretest. Such subjective recalibration of response options invalidates change measurement based on pretest and posttest scores, because measurements at both occasions are normed by different behavioral anchors. As a result, observed change scores may be high even though actual change is small, and vice versa. This type of change is known as *beta change* (Golembiewski et al., 1976).

The second type of change between pretest and posttest measures is called *gamma change* (Golembiewski et al., 1976), and occurs when respondents’ fundamental understanding and definition of a latent attribute changes between measurement occasions. For example, respondents may perceive symptoms of distress as an indication of anxiety at pretest but the therapy they undergo may have focused on recognizing different types of stressors, thus leading the measurement away from anxiety at posttest. Gamma change can hinder meaningful change assessment, because pretest and posttest scores represent conceptually different latent attributes. Hence, for valid use of outcome measures in psychotherapy it is important that both beta and gamma change are ruled out, so that observed-score change only reflects real change. Golembiewski et al. (1976) use the terminology of alpha change to identify real change.

In this study, we investigated possible beta and gamma change in the Dutch version of the Outcome Questionnaire-45 (OQ-45; Lambert et al., 1996; De Jong et al., 2007). For this purpose, we used a combination of factor analysis (FA) and item response theory (IRT). OQ-45 is a widely used self-report questionnaire for monitoring patient functioning throughout treatment in three different functional domains (Hatfield & Ogles, 2004). These functional domains are related to the symptoms of distress experienced on intrapersonal, interpersonal and societal levels. However, only when OQ-45 measurements are invariant over time can observed change on the OQ-45 be attributed to real change in these functional domains. Therefore, we aimed at investigating temporal measurement invariance in the Dutch OQ-45 by answering the following questions:

1. Is there evidence of beta change in the Dutch OQ-45 over time and if so, what are the consequences for practical change assessment?
2. Is there evidence of gamma change in the Dutch OQ-45 over time and if so, what are the consequences practical for change assessment?

In quality of life research, occurrence of beta or gamma change is interpreted as evidence of response shift (Howard et al., 1979; McPhail, Comans, & Haines, 2010; Nieuwkerk, Tollenaar, Oort, & Sprangers, 2007). Beta and gamma change have to be assessed sequentially, that is, first, one has to ascertain that the same latent attribute is being measured at all measurement occasions before proceeding to investigating possible item recalibration (Meade, Lautenschlager, & Hecht, 2005). Therefore, we first concentrate on gamma change and then on beta change.

5.2 Method

Participants and Data

A secondary data analysis was conducted using data from $N = 540$ outpatients (De Jong et al., 2007). Data were collected at three treatment departments within two medium-sized mental healthcare institutions in the Netherlands. A wide range of psychiatric disorders are treated at these institutions, including disorders related to mood, anxiety, adjustment and personality. The patients in the sample all underwent therapy and on average completed the OQ-45 3.78 times (min: once, max: 13 times, median: 3 times) throughout treatment. As pretest and posttest scores we used data from the first administration and the very last

administration, respectively. Patients who were administered OQ-45 only once were excluded from the analyses, resulting in a final sample of 412 patients.

The Outcome Questionnaire-45 (OQ-45)

The OQ-45 (Lambert et al., 1996) contains 45 Likert items with response options ranging from 0 (*never*) to 4 (*almost always*). Together the items comprise three subscales, which are the Symptom Distress (SD; 25 items) subscale, which taps symptoms of the most common types of psychological distress encountered in practice such as depression and anxiety; the Interpersonal Relations (IR; 11 items) subscale, which measures problems encountered in interpersonal relations; and the Social Role (SR; 9 items) subscale, which taps distress on a broader social level including distress encountered at work, during education, and leisure activities.

Two remarks with respect to the OQ-45 are in order. First, the hypothesized three-factor structure of OQ-45 proposed by Lambert and colleagues (1996) was found to have a poor fit to data (e.g., Beretvas & Kearny, 2003; Chapman, 2003; Kim, Beretvas, & Sherry, 2010; Mueller, Lambert, & Burlingame, 1998). In particular, in the Dutch OQ-45, De Jong et al. (2007) have identified an additional subscale containing twelve items from the SD subscale which measure symptoms of distress related exclusively to anxiety and its physical manifestations. The authors have named this subscale Anger and Somatic Distress (ASD) subscale. However, the clinical relevance of ASD as a separate scale of functioning is not yet evident (e.g., Jabrayilov, Emons, De Jong, & Sijtsma, 2015). Therefore, we included both the hypothesized factorial structure and the empirical structure resulting from our sample in the analysis.

Second, previous studies (Conijn, Emons, De Jong, & Sijtsma, 2015; De Jong et al., 2007; Jabrayilov et al., 2015) with respect to the psychometric properties of the Dutch OQ-45 revealed four items (i.e., items 11, 12, 26, & 32), which were problematic either because of poor fit with the rest of the items in the corresponding subscales or unobserved response categories (zero response frequency in the sample, suggesting irrelevant response categories and possibly affecting responses frequencies in the other categories). Therefore, consistent with the previous studies, these four items were excluded from the analyses to avoid computational problems. After the exclusion of the problematic items, 24 items remained in the SD, 10 in the IR and 7 in the SR subscales.

Data Analysis Strategy

Gamma change. In general, to assess gamma change one has to investigate whether the number of factors has changed or whether the pattern of fixed and free factor loadings for a fixed number of factors has changed from pretest to posttest (Oort, 2005; Schaubroeck & Green, 1989; Schmitt, 1982). With respect to the number of factors, based on previous studies (De Jong et al. 2007; Lambert et al., 1996) and preliminary exploratory factor analyses of our sample data, we compared the fit of three- and four-factor models at both pretest and posttest. The most parsimonious model that adequately fits the data was retained for further analysis. Gamma change was then assessed by comparing the patterns of free and fixed loadings and cross loadings between pretest and posttest in the three-factor model; that is, we tested for so called configural invariance (Widaman, Ferrer, & Conger, 2010). Gamma change was inferred when either (1) a particular item had the highest loading on different factors at pretest and posttest or (2) the number of factors on which the items had substantive loadings changed across pretest and posttest. All factor models were fitted on the polychoric correlation matrix, using MPlus5.0 (Muthén & Muthén, 1998-2001) and weighted least squares means-adjusted (WLSM) estimation. Factor analysis of polychoric correlation matrices avoids finding spurious factors in an exploratory factor analysis; these so called difficulty factors arise when the item-score distributions are skewed (Embretson & Reise, 2000).

Assessing gamma change with IRT methodology is also possible, but usually avoided because it requires the use of complex multidimensional IRT models for which estimation problems may easily occur (e.g., Meade et al., 2005; technically, factor analysis only uses univariate and bivariate data statistics, whereas IRT is based on full-information methods that take higher-order associations into account). Therefore, we concentrated on gamma change assessment using confirmatory factor analysis of polychoric correlation matrices (i.e., limited information item-factor analysis; Forero & Maydeu-Olivares, 2009).

Beta change. Beta change was assessed within the framework of IRT. We used the graded response model (GRM; Samejima, 1969), which is suitable for modeling data obtained by means of Likert items. Let θ denote the latent variable. The GRM assumes unidimensionality, local independence, and a non-linear, logistic (i.e., S-shaped) relationship between θ and the cumulative response probabilities. In particular, for each item this relationship is defined by one slope parameter (a) and M threshold (b) parameters, where M

equals the number of response categories minus 1; that is, for a Likert item, $M = 4$ (the reason is that the probability of having a score of at least 0, that is, any score, equals 1, which is a trivial result). The slope parameter expresses how well an item distinguishes low and high θ values and thus how strongly observed scores are associated with the latent variable. The threshold parameter b_m ($m = 1, \ldots, 4$ for OQ-45 Likert items) denotes the location on the θ-scale where the probability of obtaining score *m* or higher passes .50. Beta change amounts to change in the item parameters, either a, b, or both, between pretest and posttest, provided that items are calibrated on the same scale at pretest and posttest. Unidimensionality and local independence were evaluated using the residual correlations under the 1-factor model.

For testing beta change, we used likelihood-ratio tests (LRT; e.g., Lindgren, 1993) that are implemented in FlexMIRT (Houts & Cai, 2013) to test whether beta change under the graded response model was significant. In order to examine beta change, we tested equality of GRM item parameters between pretest and posttest using the LRT (Thissen, 2001). The LRT compares the likelihood of two nested models, one model that assumes that both the a and b parameters are equal at pretest and posttest (i.e., restricted model of no beta change) and one in which the a and b parameters for one or more items are freely estimated at pretest and posttest (i.e., the general model). A significant LRT suggests that fit of the restricted model is significantly worse relative to the general model, thus indicating that either the slopes or the thresholds changed from pretest to posttest.

Beta change can also be assessed by means of factor analysis. It is tested whether factor intercepts and/or factor loadings changed between pretest and posttest (e.g., Schmitt, 1982; Taris, Bok, & Meijer, 1998). Factor loadings are conceptually equivalent to slope (a) parameters in IRT. However, the interpretation of the item intercept quantified in factor analysis as a factor loading is somewhat different from the interpretation of the b parameter in IRT. The intercept in a factor analysis can be conceived as the overall item difficulty, whereas the b parameter defines the popularity of each category. In practice, item intercepts in factor analysis are rarely utilized for assessing beta change (Meade et al., 2005). More importantly, IRT is better able to exhibit subtle forms of beta change when violations of measurement invariance pertain only to some categories. Such beta changes may not be visible as changes in intercept in factor models, thus as changes in the average item difficulty.

5.3 Results

Gamma Change

Comparison of the three- and the four-factor models without restrictions on the loadings, showed only minor differences in model fit, both at pretest and posttest (Table 1). Moreover, according to the CFI and TLI (both above .95) the three-factor model had acceptable fit, and according to the RMSEA the three-factor model had moderate fit. These results suggest that a three-factor model is an adequate description of the data structure at both time points. Therefore, we proceeded with the three-factor model.

Table 1. *Model Fit Statistics for Confirmatory Factor Analysis.*

Model	Model Fit Statistics		
	RMSEA	CFI	TLI
		At Pretest	
3F-unrestriced	.086	.952	.944
4F-unrestricted	.079	.963	.954
3F-restricted	.081	.956	.951
3F-Lambert	.121	.897	.891
		At Posttest	
3F-unrestriced	.097	.970	.965
4F-unrestricted	.089	.976	.970
3F-restricted	.090	.973	.970
3F-Lambert	.144	.927	.923

To compare the pattern of factor loadings under the three-factor model between pretest and posttest, we used the three-factor model in which the items were allowed to load on all three factors as the baseline model. However, because of the small sample size relative to the number of estimated parameters and the many cross loadings, the factorial solution could be unstable, rendering its generalizability limited. Therefore, we fitted the three-factor model restricting all cross-loadings that were non-significant at both pretest and posttest restricted to zero, and such that for each factor there was at least one item that loaded only on that factor and not on the other factors. These items were used to identify the scale. The resulting model fitted well. The pattern of factor loadings that emerged in the restricted three-factor model was different from the

original three-factor model proposed by Lambert et al. (1996, 2004). Their three-factor model was also fitted to the data, but this model showed poor fit both at pretest and posttest (TLI and CFI < .95 and RMSEA > .10 at both pretest and posttest). To avoid drawing conclusions from a poorly fitting model, we proceeded with the restricted three-factor model that emerged in the current sample.

Closer inspection of the factor-loading pattern under the restricted three-factor model (Table 2) showed a consistent configural pattern of low and high loadings at pretest and posttest. For the items 3, 5, and 6 the results were ambiguous. These three items loaded on two factors, but the factor on which the items loaded highest differed between pretest and posttest. However, loadings were small, and they differed significantly at the 5% significance level (two-tailed). The standardized loadings on the posttest were generally a little higher, which may be explained by an increase in the factor variance at posttest due to inter-individual differences with respect to change after therapy. To conclude, the results suggest that even though the loadings were unequal (indicating possible beta change), the *pattern* of cross-loadings was comparable between pretest and posttest. Hence, in the Dutch OQ-45 gamma change over time is absent. However, the factorial structure is inconsistent with theoretical expectations derived from Lambert et al. (1996, 2004), both at pretest and posttest.

Beta Change

For beta change analysis, we adopted the composition of the SD, IR, and SR subscales (see Lambert et al., 1996, 2004; De Jong et al., 2007) to make sure that the results are consistent with the practical use of the OQ-45. Previous IRT analyses of the same subscales (Jabrayilov et al., 2015) showed adequate fit of the GRM. In particular, inspection of the residual correlations under the one-factor model revealed few values in excess of .15 (Morizot, Ainsworth, & Hayes, 2007), indicating few local dependencies. Local dependencies may hinder effective IRT modeling, because they may inflate the estimated a parameters. Therefore, for locally dependent item pairs it was tested whether α parameter estimates were significantly inflated using the Jackknife Slope Index (JSI; Edwards & Cai, 2011). For each subscale, none of the JSIs was significant at the 5% significance level. This means that we found no evidence that local dependencies bias the a estimates. Therefore, we proceeded assessing beta change at the subscale level, assuming unidimensionality.

Table 2. *Factor Loadings for the Confirmatory Three-Factor Model.*

		Pretest			Posttest		
# item	Hyp. 3F	F1	F2	F3	F1	F2	F3
1	F1	**.581**	-.166	.193	**.658**	-.173	.265
2	F2		**.543**	.135		**.645**	.041
3	F2	.267	**.353**	.097	**.425**	.369	.056
4	F3	-.065	.295	**.611**	.023	.124	**.704**
5	F2	**.404**	.358		.400	**.462**	
6	F2	**.324**	.307	.229	.243	**.443**	.169
7	F1	**.269**		.149	**.342**		.308
8	F2	**.362**	**.388**	.096	**.302**	**.397**	.133
9	F2	.179	**.660**		.168	**.712**	
10	F2	-.046	**.785**	-0.16	-.087	**.840**	-.105
13	F2	**.683**	.215		**.715**	.187	
14	F3	-.179		**.507**	-.248		**.595**
15	F2	**.478**	.445		**.523**	.468	
16	F1		**.357**			**.463**	
17	F1	**.319**		.140	**.323**		.276
18	F1	**.451**	.294		**.491**	.389	
19	F1	**.262**		.208	**.391**		.324
20*	F1	**.732**			**.784**		
21	F3	**.512**	.279		**.679**	.113	
22	F2		**.545**	.182		**.647**	.103
23	F2	.349	**.480**		.339	**.553**	
24	F2	**.653**	.266	-.172	**.669**	.241	-0.93
25	F2		**.673**			**.723**	
27	F2		**.421**			**.518**	
28	F3		.087	**.247**		.123	**.318**
29	F2	-.144	**.630**	.071	.002	**.669**	.090
30	F1	**.604**		.271	**.593**		.263
31	F2	**.735**	.195		**.776**	.135	
33	F2		**.632**			**.751**	
34	F2		**.412**	.051		**.476**	.108
35	F2		**.526**			**.568**	
36*	F2		**.766**			**.770**	
37	F1	**.619**	-.174	.203	**.759**	-.176	.205
38	F3		.206	**.626**		.126	**.763**
39*	F3			**.787**			**.737**
40	F2	.286	**.442**		.210	**.550**	
41	F2		**.528**			**.559**	
42	F2	.336	**.580**		.345	**.615**	
43	F1	**.774**	-.103	.242	**.866**	-.213	.260
44	F3	.167		**.711**	.182		**.668**
45	F2		**.435**			**.568**	

Note. * Used for identification in restricted model. Non-significant cross-loadings were restricted to zero; For each item the largest loadings at pretest and posttest are printed in boldface; unreported cross-loadings were fixed to zero in the model. Hyp. 3F = hypothesized three-factor model of Lambert et al. (2004).

The LRT for testing beta change across time requires a subset of time-invariant items, also known as anchor set, which can be used to account for real change in functioning at pretest and posttest (Orlando-Edelen, Thissen, Teresi, Kleinman, & Ocepek-Welikson, 2006).

A commonly used strategy to empirically select the anchor set is scale purification (Gonzales-Betanzos & Abad, 2012). The purification procedure first takes the whole set of items as the initial anchor set. Each item in the initial anchor set is tested for significant beta change, using the other items as the anchor items. The item showing the highest beta change is removed from the anchor set, thus producing a new initial anchor containing one item fewer than the previous set. This procedure is repeated until a final set of anchor items is found without items showing significant beta change. To avoid inflated Type I error rate, in each iteration we used a Bonferroni corrected significance level of $.05/k$, where k represents the number of tested items.

The scale purification process revealed two items with potential beta change over time. These were items 38 ("I feel that I am not doing well at work/school") from the SR subscale, and item 42 ("I feel blue") from the SD subscale. Final LRT of these items using purified anchors confirmed significant beta change in either as or bs: $\chi^2(5) = 21.8, p < .01$ for item 38, and $\chi^2(5) = 22.2, p < .01$ for item 42. For item 38, beta change was caused by a change in both the as or bs, whereas for item 42, only the bs were significantly different between pretest and posttest. Table 3 shows the estimated item parameters for these items at pretest and posttest.

Table 3. *Estimated Item Parameters for the Graded Response Model at Pretest and Posttest for Items 38 and 42.*

Measurement Occasion	Estimated Item Parameters				
	a	b_1	b_2	b_3	b_4
	I feel that I am not doing well at work/school (item 38)				
Pretest	2.15	-0.78	0.05	0.90	1.91
Posttest	3.44	-0.82	0.16	1.05	1.83
	I feel blue (item 42)				
Pretest	2.83	-1.23	-0.68	0.42	1.72
Posttest	3.16	-1.55	-0.53	0.65	1.69

To assess the practical impact of beta change on measurement, for each item we compared the relationship between the expected item score and θ (Figure 1) and expected

total score and θ (Figure 2) between pretest and posttest. The figures suggested that the impact of beta change on practical measurement was minimal. Conditional on θ, the largest difference between the expected items scores at pretest and posttest was 0.2. This means that on average beta change explained at most a change of 0.2 item-score units. Given that the items are scored on a 5-point scale, we consider a bias of 0.2 to be practically unimportant. Moreover, the effect beta change in items 38 and 42 had on the expected total score was negligible as the curves representing pretest and posttest total scores were indistinguishable. Therefore, we concluded that even though items 38 and 42 showed significant beta change between pretest and posttest, the impact of beta change on practical change assessment in the Dutch OQ-45 was negligible.

5.4 Discussion

Response shift involving gamma change or beta change, is considered an important threat to the validity of change scores obtained in pretest-posttest designs (e.g., Howard et al., 1979; McPhail, Comans, & Haines, 2010; Nieuwkerk, Tollenaar, Oort, & Sprangers, 2007). However, a definitive conclusion regarding the prevalence of response shift and its impact on change assessment has not been drawn (Schwarts et al., 2006). Our study provides evidence that the Dutch OQ-45 can be used safely in change assessment based on pretest and posttest scores despite the beta change in two items.

Some of the issues to consider with respect to our study are the following. First, results for gamma and beta change were based on measurements from the very first session and the last time the patients completed the OQ-45. Hence, we did not include data on interim administrations. However, given that we used posttest data that were most distant in time from pretest, and given that we did not find gamma or important beta change between these measurements, we hypothesize that these results also generalize to the other administrations. Second, the LRT for beta change with scale purification assumed that there is also a set of items that do not show beta change. However, when all items show equal amounts of beta change, the beta change is absorbed in the latent variable distribution and the purification process will not find potentially biased items.

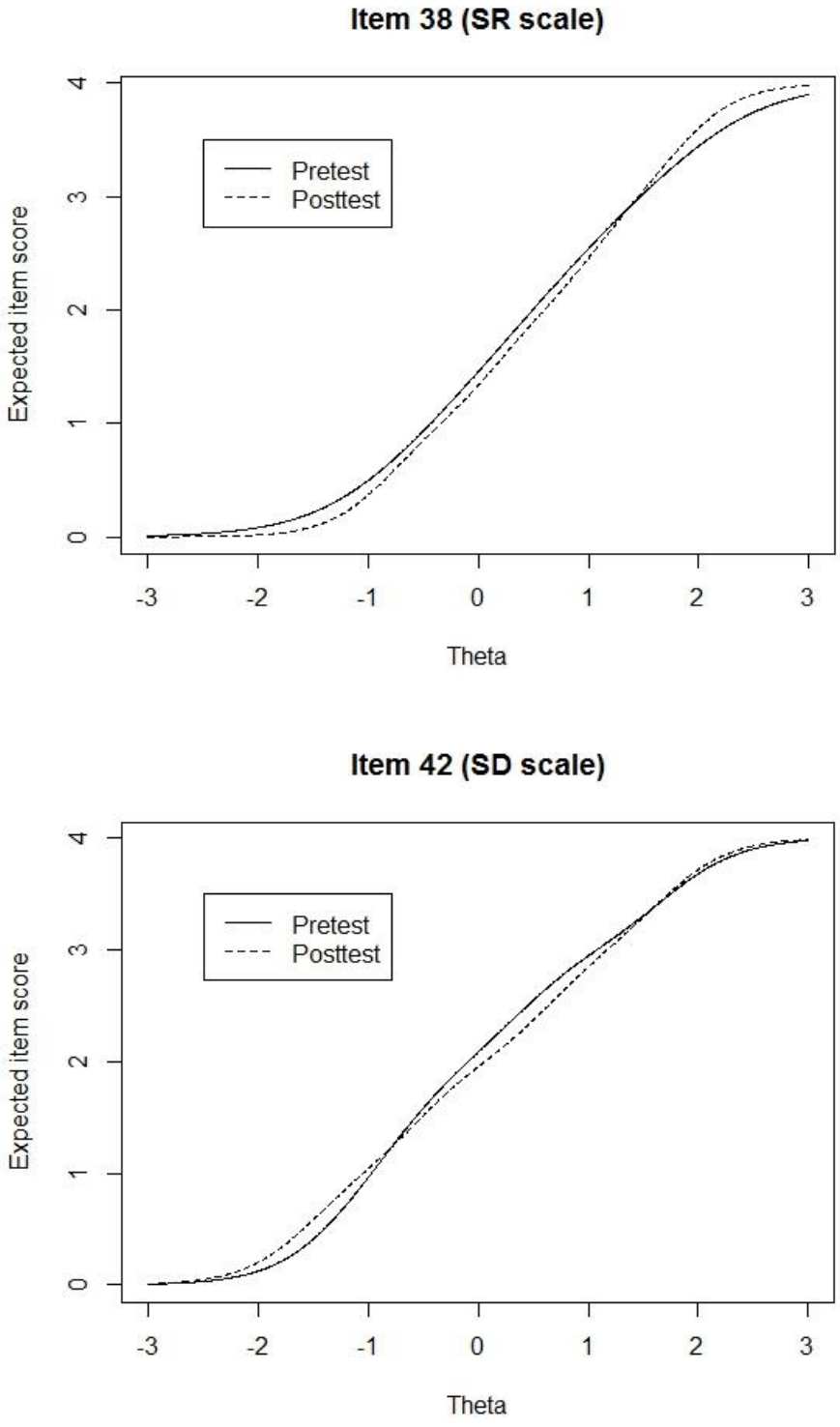


Figure 1. *Expected item scores for items 38 and 42 as a function of θ. Note. SR: Social Role; SD: Symptom Distress.*

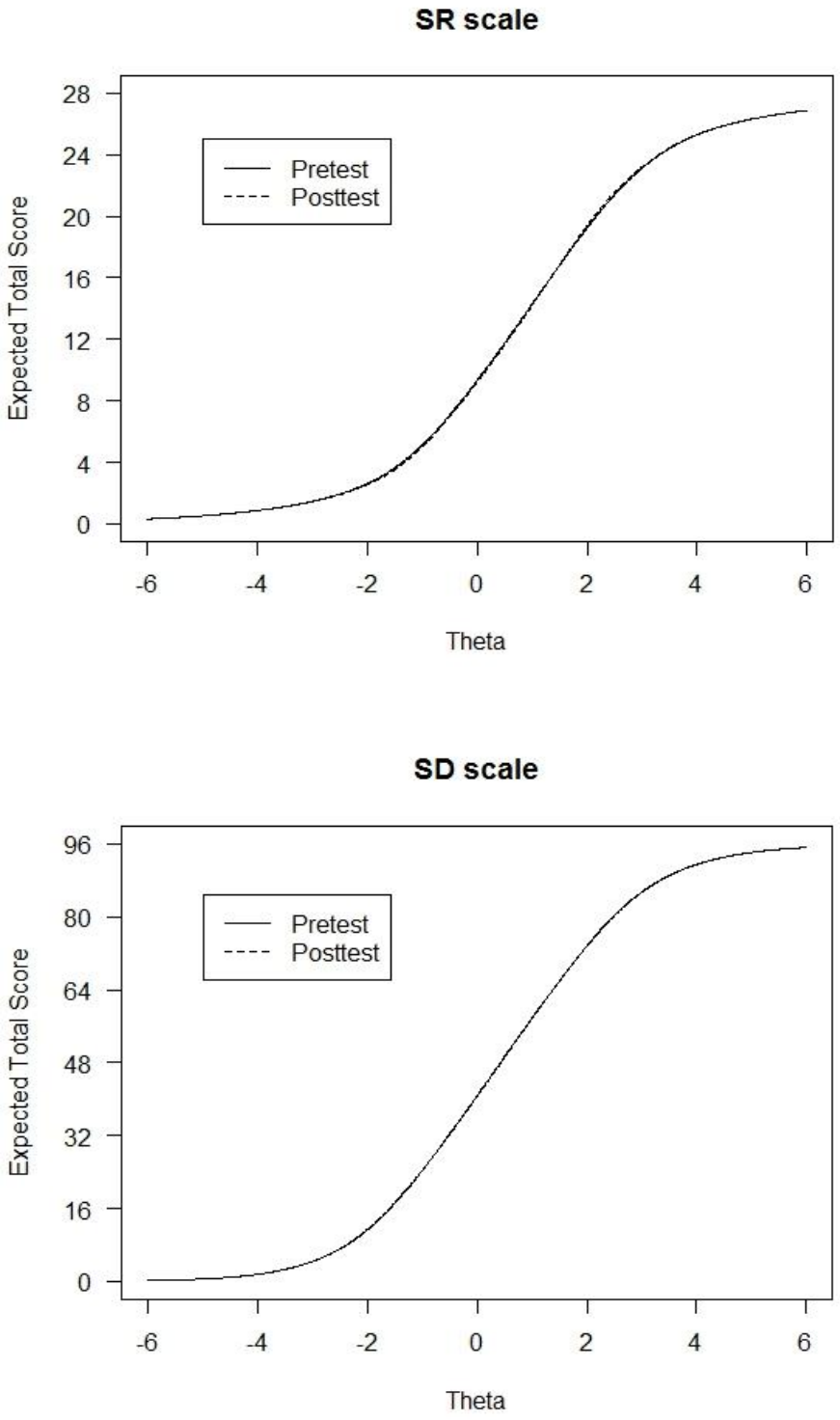


Figure 2. *Expected total scores for the SR and SD scales as a function of θ. Note. SR: Social Role; SD: Symptom Distress.*

Hence, when beta change is equal across all items, beta change goes undetected using the LRT with purification approach.Uniform beta change across all items may appear unlikely, but future research may focus on alternative psychometric methods for detecting uniform beta change to rule out this possibility.

We did not find gamma change exhibited by a factor structure that changed from pretest to posttest. However, the factor structure found differed from the hypothesized three-factor solution Lambert et al. (1994, 2004) proposed. This result is consistent with previous studies on the factor structure of the OQ-45 (Beretvas & Kearny, 2003; Chapman, 2003; Kim, Beretvas, & Sherry, 2010; Mueller, Lambert, & Burlingame, 1998), which also failed to support the hypothesized three-factor structure. It is not clear what explains these inconsistencies, but it seems that different clinical populations entertain different conceptualizations of items (Kim et al. 2010), and conceptualization of items may even vary across individual persons. For example, item 21 (“I enjoy my spare time”) was assigned to the SR scale, but we found a high loading on the factor related to SD. We believe this is not very surprising, because failing to enjoy spare time in the general population may be due to poor social relationships, but in clinical patients suffering from depressive thoughts failure may be driven by distress. To conclude, gamma change analyses suggested that the same attribute is being measured at pretest and posttest, but which attribute is being measured is unclear, and also how well the attribute generalizes to other populations.

For the beta change analysis we used the GRM. In spite of the ambiguous factorial structure and the many cross loadings, the GRM fitted the subscales surprisingly well and all items in the scale contributed to reliable measurement of the underlying factor. The adequate fit can be explained by the high correlations between the factors and the many cross loadings causing items to be informative about the underlying attribute even though factor analysis assigns the item to a different scale. Another issue when using LRT in IRT is the assumption of local independence. The procedure treats pretest and posttest measures as independent random samples from different populations. Hence, the procedure assumes zero-correlated measurement errors at the individual level. This is a highly restrictive assumption, but we notice that zero-correlated errors are also assumed when testing individual change for significance using the reliable change index (Jacobson & Truax, 1991). For assessing gamma change, we estimated the models separately at pretest and posttest, such that correlated errors at the item level, if any, do not play a role.

This study focused on evidence of beta or gamma change at the group level. However, group-level indicators may hide important information about beta or gamma change within individual patients in these groups. This means that absence of evidence of response shift at the group level leaves open the possibility that individual patients changed their responses independent of their health change. Future research may focus on methods for detecting response shifts within individuals in pretest-posttest designs. One approach could be person-fit analysis (Meijer & Sijtsma, 2001), which aims at detecting individuals whose response pattern is unlikely given the measurement model. Person fit-analyses has been applied successfully to explain cross-sectional differences in aberrant responding behavior to the Dutch OQ-45 (Conijn et al., 2015), and time-dependent differences in aberrant response behavior on the Spielberger's Stait and Trait Anxiety Inventory (Conijn, Emons, Van Assen, Pedersen, & Sijtsma, 2013). This line of research may be continued by considering dedicated person-fit methods for detecting response shift within individuals. Such person-fit procedures may warn the therapist against validity failure of self-reported individual outcomes.

To our knowledge, this study was the first attempt to assess temporal measurement invariance in measurement by means of the Dutch OQ-45 using a sample of outpatients. Even though we did not find evidence of response shift, we think it is premature to draw general conclusions with respect to the absence of beta or gamma change in measurement using the OQ-45. More studies are needed to deepen our understanding of measurement invariance in OQ-45. Also, future studies should take the limitations of our study into account and re-assess our recommendations.

Summary

In clinical settings, assessment of individual persons' change based on at least two test scores is an important procedure to evaluate improvement or deterioration as a result of therapy. The psychometrics of change assessment is the topic of this Ph.D. thesis.

Because of its item-level approach to test scoring, item response theory (IRT) is generally believed to be superior in assessing individual change than traditional methods from classical test theory (CTT). However, IRT requires more technical knowledge and, as a result, elaborate calculations based on IRT can be daunting for those lacking the necessary background to apply it to real-data analysis. Moreover, little research is available that supports the exact advantages and disadvantages of using one methodology over the other in change assessment. This issue is further complicated by the availability of several different estimation methods within the IRT framework, each of which may lead to different conclusions about change in individual patients. Therefore, deciding which method to use for change assessment can be too difficult for researchers lacking the necessary background in psychometrics.

A considerable part of this thesis was dedicated to studying the differences between various test-scoring methods with respect to change assessment as well as demonstration of the practical application of IRT to change assessment in clinical settings. In the first study (Chapter 2), we compared the accuracy of three widely-used estimation methods in IRT (i.e., maximum likelihood (ML), expected aposteiori (EAP) and weighted maximum likelihood (WML)) with respect to the detection of statistical significance of individual change scores based on the JT method. The results of our simulation study showed that, even though the differences between the three estimation methods were small, for shorter tests (at most 10 items) WML was the most accurate. For longer tests (at least 20 items), all three methods performed equally well. From this study we concluded that it is better to use (1) WML for short tests and EAP for longer tests because EAP is computationally less intensive, and (2) longer tests in general as they are more accurate in detecting statistical significance of individual change scores irrespective of which estimation methods one uses.

In the second and third studies, we compared CTT and IRT with respect to individual change assessment. These studies were based on simulated and real data, respectively. The results of Study 2 (Chapter 3) revealed that, although the differences between the two methods were small, compared to CTT, IRT detected individual change based on the JT method better, provided the tests consist of at least 20 items. However, compared to IRT, for shorter tests CTT was better at detecting change in individuals. In general, the results of Study 2 showed that short tests (at most ten items) are not optimal for change assessment irrespective of whether one uses CTT or IRT. Therefore, we concluded that for change assessment tests consisting of at least 20 items and scored with IRT should be preferred

In Study 3 (Chapter 4), we found that, in addition to CTT or IRT, the method by which one judges the significance of change also affects change assessment. We found that IRT is more likely to classify patients as having significantly changed if one uses the JT method. However, the differences between the CTT and IRT were not clear when we used another method for assessing change, that is, the minimal clinically important difference (MCID). Therefore, instead of recommending the exclusive use of IRT or CTT for individual change assessment, we concluded that each method has its own advantages and disadvantages depending on factors such as the individual change assessment method (i.e., JT, MCID) and the test length.

In the last study (Chapter 5), we used CTT and IRT methodology in combination to study temporal measurement invariance in the Dutch version of the Outcome Questionnaire-45 (OQ-45). The OQ-45 is a frequently used instrument for assessing change in clinical settings. Temporal measurement invariance is an important prerequisite for change assessment based on pretest and posttest scores. To ascertain temporal measurement invariance, one has to assure that the factorial structure as well as the interpretation of the item response options is the same at pretest and posttest. Our results showed that, despite the small violations of measurement invariance over time, change assessment by means of the Dutch OQ-45 based on pretest and posttest yields valid results.

References

Alvares, K. M., & Hulin, C. L. (1972). Two explanations of temporal changes in ability-skill relationships: A literature review and theoretical analysis. *Human Factors*, *14*, 295–308.

Arthur, W., & Day, D. W. (1994). Development of a short-form of the Raven Advanced Progressive Matrices. *Educational and Psychological Measurement, 54*, 394-403.

Baker, F. B., & Kim, S-H. (2004). *Item response theory. Parameter estimation techniques* (second edition). New York, NY: Dekker Media.

Bauer, S., Lambert, M., & Nielsen, S. L. (2004). Clinical significance methods: A comparison of statistical techniques. *Journal of Personality Assessment, 82,* 60-70.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry, 4*, 561–571.

Beretvas, S. N., & Kearney, L. K. (2003). *A shortened form of the Outcome Questionnaire: A validation of scores across groups (A research report of the Research Consortium of Counseling and Psychological Services for Higher Education)*. Austin: University of Texas at Austin, Counseling and Mental Health Center.

Bock, R. D., & Zimowski, M. F. (1997). Multiple group IRT. In W. J. van der Linden & R. K. Hambleton (eds). *Handbook of modern item response theory.* New York, NY: Springer.

Boswell, J. F., Kraus, D. R., Miller, S. D., & Lambert, M. J. (2013). Implementing routine outcome monitoring in clinical practice: Benefits, challenges, and solutions. *Psychotherapy Research, 25*(1), 6-19.

Breetvelt, I. S., & Van Dam, F. S. A. M. (1991). Underreporting by cancer patients: the case of response-shift. *Social Science Medicine*, *32*, 981-987.

Brouwer, D. (2013). Modern psychometric perspectives on the evaluation of clinical scales (Doctoral dissertation, University of Groningen). Retrieved from https://www.rug.nl/research/portal/files/2404658/Dissertation_DBrouwer_2013-1.pdf

Carlier, M., & Roubertoux, P. (2010). Genetics and cognition: The impact for psychologists in applied settings. *European Psychologist*, *15*, 49– 57.

Chapman, J. E. (2003). *Reliability and validity of the progress questionnaire: An adaptation of the Outcome Questionnaire*. Philadelphia, PA: Drexel University.

References

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale NJ: Lawrence Erlbaum Associates.

Conijn, J. M., Emons, W. H. M., De Jong, K., & Sijtsma, K. (2015). Detecting and explaining aberrant responding to the Outcome Questionnaire–45. *Assessment 22*(4), 513-524.

Conijn, J. M., Emons, W. H. M., Van Assen, M. A. L. M., Pedersen, S. S., & Sijtsma, K. (2013). Explanatory, multilevel person-fit analysis of response consistency on the Spielberger State-Trait Anxiety Inventory. *Multivariate Behavioral Research, 48*(5), 692-718.

Copay, A. G., Subach, B. R., Glassman, S. D., Polly, D. W., & Schuler, T. C. (2007). Understanding the minimum clinically important difference: a review of concepts and methods. *The Spine Journal, 7,* 541-546.

Crowder, B., & Michael, W. B. (1991). The development and validation of a short form of a multidimensional self-concept measure for high technology employees. *Educational and Psychological Measurement, 51*, 447-454.

De Cock, E. S., Emons, W. H. M., Nefs, G., Victor, J. M. P., & Pouwer, F. (2011). Dimensionality and scale properties of the Edinburgh Depression Scale (EDS) in patients with type 2 diabetes mellitus: the DiaDDzoB study. *BMC Psychiatry*, 11:141.

De Jong, K., Nugter, M. A., Polak, M. G., Wagenborg, J. E. A., Spinhoven, P., & Heiser, W. J. (2007). The Outcome Questionnaire (OQ-45) in a Dutch populations: A cross-cultural validation. *Clinical Psychology and Psychotherapy, 14*, 288–301.

De Jong, K., Nugter, M. A., Lambert, M. J., & Burlingame, G. M. (2009). *Handleiding voor afname en scoring van de Outcome Questionnaire (OQ-45).* [Manual for administration and scoring of the Outcome Questionnaire (OQ-45)]. Salt Lake City, UT: OQ Measures LLC

Doucette, A., & Wolf, A. W. (2009). Questioning the measurement precision of psychotherapy research, *Psychotherapy Research, 19*, 374-389.

Drasgow, F., Levine, M. V., Tsien, S., Williams, B., & Mead, A. D. (1995). Fitting polytomous item response theory models to multiple-choice tests. *Applied Psychological Measurement, 19,* 143-165.

Edwards, M. C., & Cai, L. (2011). *A new procedure for detecting departures from local independence in item response models*. Paper presented at the annual meeting of American Psychological Association, Washington, D.C. Retrieved from http://faculty.psy.ohio-state.edu/edwards/documents/APA8.2.11.pdf

References

Embretson, S. E. (2006). The continued search for nonarbitrary metrics in psychology. *American Psychologist, 61*, 50-55.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah NJ; Lawrence Erlbaum Associates.

Emons, W. H. M., Sijtsma, K., & Meijer, R. R. (2007). On the consistency of individual classification using short scales. *Psychological Methods, 12, 105-120.*

Finkelman, M. D., Weiss, D. J., & Kim-Kang, G. (2010). Item selection and hypothesis testing for the adaptive measurement of change. *Applied Psychological Measurement, 34*, 238-254.

Forero, C. G., Maydeu-Olivares, A., & Gallardo-Pujol, D. (2009). Factor analysis with ordinal indicators: A Monte Carlo study comparing DWLS and ULS estimation. *Structural Equation Modeling*, *16*(4), 625-641.

Golembiewski, R. T., Billingsley, K., & Yeager, S. (1976). Measuring change and persistence in human affairs: Types of change generated by OD designs. *Journal of Applied Behavioral Science*, *12*, 133–157.

González-Betanzos, F., & Abad, F. J. (2012). The effects of purification and the evaluation of differential item functioning with the likelihood ratio test. *Methodology*, *8*(4), 134-145.

Gosling, S. D., Rentfrow, P. J., & Swann, Jr., W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality, 37*, 504-528.

Hambleton, R. K., & Swaminathan, H. (1985). *Item Response Theory. Principles and Applications*. New York. NY: Springer.

Hatfield, D. R., & Ogles, B. M. (2004). The use of outcome measures by psychologists in clinical practice. *Professional Psychology—Research and Practice*, *35*(5), 485–491.

Hays, R. D., Brodsky, M., Johnston, M. F., Spritzer, K. L., & Hui, K-K. (2005). Evaluating the statistical significance of health-related quality of life change in individual patients. *Evaluation & The Health Professions, 28,* 160-171.

Henry, R. A., & Hulin, C. L. (1987). Stability of skilled performance across time: Some generalizations and limitations on utilities. *Journal of Applied Psychology*, *72*, 457–462.

Hiller, W., Schindler, A. C., & Lambert, M. J. (2011). Defining response and remission in psychotherapy research: A comparison of the RCI and the method of percent improvement. *Psychotherapy Research, 22, 1-11.*

Hill, C. D., Edwards, M. C., Thissen, D., Langer, M. M., Wirth, R. J., Burwinkle, T. M., & Varni, J. W. (2007). Practical issues in the application of item response theory [Supplemental material]. *Medical Care, 45, 39-47.*

Hooker, G., Finelman, M., & Schwartzman, A. (2009). Paradoxical results in multidimensional item response theory. *Psychometrika, 74,* 419-442.

Houts, C. R., & Cai, L. (2013). *flexMIRTR user's manual version 2: Flexible multilevel multidimensional item analysis and test scoring.* Chapel Hill, NC: Vector Psychometric Group.

Howard, G. S., Ralph, K. M., Gulanick, N. A., Maxwell, S. E., Nance, S. W., & Gerber, S. K. (1979). Internal invalidity in pre-test-posttest self-report evaluations and a re-evaluation of retrospective pre-tests**.** *Applied Psychological Measurement*, *3*, 1–23.

IBM Corporation. (2013). *IBM SPSS Statistics for Windows*, *Version 22.0*. Armonk, NY: IBM Corp.

Jabrayilov, R., Emons, W. H. M., & Sijtsma, K. (2014). *Comparison of three latent trait estimation methods in reliable change assessment.* Manuscript submitted for publication.

Jabrayilov, R., Emons, W. H. M., & Sijtsma, K. (2015). *Change assessment using IRT: An illustration and comparison with CTT-based change assessment*. Manuscript submitted for publication.

Jabrayilov, R., Emons, W. H. M., & Sijtsma, K. (2015). *Comparison of Classical Test Theory and Item Response Theory in Individual Change Assessment.* Manuscript submitted for publication.

Jacobson, N. S., Foilette, W. C., & Revenstorf, D. (1984). Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance. *Behavior Therapy, 15,* 336-352.

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy and research. *Journal of Consulting and Clinical Psychology, 59*, 12-19.

Jacobson, N. S., Roberts, L. J., Berns, S. B., & McGlinchey, J. B. (1999). Methods for defining and determining the clinical significance of treatment effects: Description, application, and alternatives. *Journal of Consulting and Clinical Psychology, 67,* 300–307.

Kazdin, A. E., & Wilson, G. T. (1978). *Evaluation of behavior therapy. Issues, evidence, and research strategies*. Cambridge: Ballinger.

References

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary Feedback Intervention Theory. *Psychological Bulletin, 119*, 254-284.

Kolen, M. J., & Brennan, R. L. (1995). *Test equating: Methods and practices*. New York: Springer.

Kruyen, P. M., Emons, W. H. M., & Sijtsma, K. (2012). Test length and decision quality in personnel selection: When is short too short? *International Journal of Testing*, *12*, 321-344.

Kruyen, P. M., Emons, W. H. M., & Sijtsma, K. (2013a). On the shortcomings of shortened tests: A literature review. *International Journal of Testing*, *13*, 223-248.

Kruyen, P. M., Emons, W. H. M., & Sijtsma, K. (2013b). Shortening the S-STAI: Consequences for research and clinical practice. *Journal of Psychosomatic Research*, *75*, 167-172.

Kruyen, P. M., Emons, W. H. M., & Sijtsma, K. (2014). Assessing individual change using short tests and questionnaires. *Applied Psychological Measurement, 38,* 201-216.

Lambert., M. J., Burlingame, G. M., Umphress, V., Hansen, N. B., Vermeersch, D. A., Clouse, G. C., & Yanchar, S. C. (1996). The reliability and validity of the outcome questionnaire. *Clinical Psychology and Psychotherapy, 3*, 249-258.

Lambert, M. J., Hansen, N. B., Umphress, V., Lunnen, K., Okiishi, J., Burlingame, G., & Reisinger, C. W. (1996). *Administration and scoring manual for the Outcome Questionnaire (OQ45.2)*. Wilmington, DE: American Professional Credentialing Services.

Lambert, M. J., Morton, J. J., Hatfield, D. R., Harmon, C., Hamilton, S., Shimokawa, K., et al. (2004). *Administration and scoring manual for the OQ-45.2 (Outcome Questionnaire)* (3rd ed.). Wilmington, DE: American Professional Credentialing Services LLC.

Lance, C. E., & Vandenberg, R. J. (Eds.) (2008). *Statistical and Methodological Myths and Urban Legends: Doctrine, Verity and Fable in Organizational and Social Sciences*. New York: Routledge.

Lindgren, W. (1993). *Statistical Theory (4th ed.)*. New York, NY: Chapman & Hall.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Lord, F. M. (1983). Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability. *Psychometrika, 48*, 233-245.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA; Addison-Wesley.

Maassen, G. H. (2004). The standard error in the Jacobson and Truax Reliable Change Index (the classical approach to the assessment of reliable change). *Journal of the International Neuropsychological Society, 10*, 888-893.

Magis, D. (2014). Accuracy of asymptotic standard errors of the maximum and weighted likelihood estimators of proficiency levels with short tests. *Applied Psychological Measurement, 38*, 105-121.

McPhail, S., Comans, T., & Haines, T. (2010). Evidence of disagreement between patient-perceived change and conventional longitudinal evaluation of change in health-related quality of life among older adults. *Clinical Rehabilitation*, *24*, 1036-1044.

Meade, A. W., Lautenschlager, G. J., & Hecht, J. E. (2005). Establishing measurement equivalence and invariance in longitudinal data with item response theory. *International Journal of Testing*, *5*, 279-300.

Meier, S. T. (2008). *Measuring change in counseling and psychotherapy.* New York, NY: The Guilford Press.

Meijer, R., & Baneke, J. J. (2004). Analyzing psychopathology items: A case for nonparametric item response theory modeling. *Psychological Methods, 9*, 354-368.

Meijer, R., & Sijtsma, K. (2001). Methodology Review: Evaluating person fit. *Applied Psychological Measurement*, *25*, 107-135.

Mellenbergh, G. J. (2011). *A conceptual introduction to psychometrics*. Den Haag, the Netherlands: Eleven International.

Meltzoff, J., & Kornreich, M. (1970). *Research in psychotherapy*. New York: Atherton.

Millsap, R. (2010). Testing measurement invariance using item response theory in longitudinal data: an introduction. *Child Development Perspectives, 4*(1), 5-9.

Mollenkopf, W. G. (1949). Variation of the standard error of measurement. *Psychometrika, 14*, 189-229.

Montgomery, S. A., & Asberg, M. (1979). A new depression scale designed to be sensitive to change. *British Journal of Psychiatry, 134,* 382–389.

Morizot, J., Ainsworth, A. T., & Reise, S. P. (2007). Toward modern psychometrics: Application of item response theory models in personality research. In R. W. Robins, R. C.

Fraley, & R. F. Krueger (Eds.). *Handbook of Research Methods in Personality Psychology* (407-423), New-York: Guilford Press.

Mueller, R. M., Lambert, M. J., & Burlingame, G. M. (1998). Construct validity of the outcome questionnaire: A confirmatory factor analysis. *Journal of Personality Assessment*, *70*, 248–262.

Mumford, M. D., Stokes, G. S., & Owens, W. A. (1990). *Patterns of life history: The ecology of human individuality*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Muthén, L. K., & Muthén, B. O. (1998-2011). *Mplus User's Guide*. *Sixth Edition*. Los Angeles, CA: Muthén & Muthén.

Nieuwkerk, P. T., Tollenaar, M. S., Oort, F. J., & Sprangers, M. A. G. (2007). Are retrospective measures of change in quality of life more valid than prospective measures? *Medical Care*, *45*, 199-205.

Norman, G. R., Sloan, J. A., & Wyrwich, K. W. (2003). Interpretation of changes in health related quality of life: The remarkable universality of half a standard deviation. *Medical Care, 41,* 582-592.

Nydick, S. W. (2012). *catIrt: An R Package for Simulating IRT-Based computerized adaptive tests. R package version 0.3-0*. http://CRAN.R-project.org/package=catIrt

Oort, F. J. (2005). Using structural equation modeling to detect response shifts and true change. *Quality of Life Research*, *14*, 587–598

Orlando-Edelen, M., Thissen, D., Teresi, J. A., Kleinman, M., & Ocepek-Welikson, K. (2006). Identification of differential item functioning using item response theory and the likelihood based model comparison approach: Applications to the Mini-Mental State Examination. *Medical Care*, *44*, 134-142.

Prieler, J. A. (2007). So wrong for so long. Changing our approach to change. *The Psychologist*, *20*, 730-732.

R Development Core Team (2014). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Reckase, M. D. (2009). *Multidimensional Item Response Theory.* New York, NY: Springer-Verlag.

Reise, S. P. (2005). Item response theory and its applications for cancer outcome measurement. In J. Lipscomb, C. C. Gotay, & C. Snyder (Eds.). *Outcome assessment in*

cancer: measures, methods, and applications. Cambridge University Press, UK: Cambridge.

Reise, S. P., & Haviland, M. G. (2005). Item response theory and the measurement of clinical change. *Journal of Personality Assessment. 84*, 228-238.

Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality in health outcomes measures. *Quality of Life Research, 16,* 19-31.

Reise, S. P., & Revicki, D. A. (2015). *Handbook of item response theory modelling: Applications to typical performance assessment*. Hove, East Sussex: Routledge.

Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology*, *5*, 27-48.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograpph*, No. 17.

Samejima, F. (1998). *Some considerations in eliminating biases in ability estimation in computerized adaptive testing*. Paper presented at the annual meeting of the American Educational Association, San Diego.

Sebille, V., Hardouin, J-B., Le Néel, T., Kubis, G., Boyer, F., Guillemin, F., et al. (2010). Methodological issues regarding power of classical test theory (CTT) and item response theory (IRT)-based approaches to comparison of patient-reported outcomes in two groups of patients – a simulation study. *BMC Medical Research Methodology,* 10:24.

Schaubroeck, J., & Green, S. G. (1989). Confirmatory factor analytic procedures for assessing change during organizational entry. *Journal of Applied Psychology*, *74*, 892–900.

Schmitt, N. (1982). The use of analysis of covariance structures to assess beta and gamma change. *Multivariate Behavioral Research*, *17*, 343–358.

Schwartz, C., Bode, R., Repucci, N., Becker, J., Sprangers, M., & Fayers, P. (2006). The clinical significance of adaptation to changing health: A meta-analysis of response shift. *Quality of Life Research*, *15*, 1533–1550.

Shimokawa, K., Lambert, M. J., & Smart, D. (2010). Enhancing treatment outcome of patients at risk of treatment failure: Meta-analytic and mega-analytic review of a psychotherapy quality assurance system. *Journal of Consulting & Clinical Psychology*, *78*, 298-311.

References

Sijtsma, K. (2012). Future of psychometrics: Ask what psychometrics can do for psychology. *Psychometrika, 77*, 4–20.

Sijtsma, K., Emons, W. H. M., Bouwmeester, S., Nyklicek, I., & Roorda, L. D. (2008). Nonparametric IRT analysis of Quality-of-Life Scales and its application to the World Health Organization Quality-of-Life Scale (WHOQOL-Bref). *Quality of Life Research, 17*, 275-290.

Sijtsma, K., & Molenaar, I. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks, CA: Sage.

Spielberger, C. D., Gorsuch, R. L., Lushene, R., Vagg, P. R., & Jacobs, G. A. (1983). *Manual for the State-Trait Anxiety Inventory*. Palo Alto, CA: Consulting Psychologists Press.

Taris, T. W., Bok, I. A., & Meijer, Z. Y. (1998). Assessing stability and change of psychometric properties of multi-item concepts across different situations: A general approach. *Journal of Psychology*, *132*, 301–316.

Thissen, D. (2001). *IRTLRDIF v.2.0b. Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning*. Chapel Hill: University of North Carolina.

Thissen, D., Pommerich, M., Billeaud, K., & Williams, V. S. L. (1995). Item response theory for score including polytomous items with ordered responses. *Applied Psychological Measurement, 19,* 39-49.

Thomas, M. L. (2011). The value of item response theory in clinical assessment: A review. *Assessment, 18,* 291-307.

Timman, R., De Jong, K., & De Neve-Enthoven, N. (2016). Cut-off scores and clinical change indices for the Dutch Outcome Questionnaire (OQ-45) in a large sample of normal and several psychotherapeutic populations, *Clinical Psychology and Psychotherapy* : 1-10.

Van der Linden, W. J. (2012). On compensation in multidimensional response modeling. *Psychometrika*, *77*, 21–30.

Van der Linden, W. J., & Hambleton, R. K. (Eds.) (1997). *Handbook of modern item response theory*. New York, NY: Springer.

Van Rijn, P. W., & Rijmen, F. (2012). *A note on explaining away and paradoxical results in multidimensional item response theory*. Princeton, NJ: Educational Testing Service, Research Report, 12-13.

Vermeersch, D. A., Lambert, M. J., & Burlingame, G. M. (2000). Outcome Questionnaire: Item sensitivity to change. *Journal of Personality Assessment, 74*(2), 242–261.

Vermeersch, D. A., Whipple, J. L., Lambert, M. J., Hawkins, E. J., Burchfield, C. M., & Okiishi, J. C. (2004). Outcome Questionnaire: Item sensitivity to changes in counseling center clients. *Journal of Counseling Psychology, 51*, 38–49.

Waller, N. G., & Reise, S. P. (2010). Measuring psychopathology with non-standard item response theory models: fitting the four-parameter model to the Minnesota Mulitphasic Personality Inventory. In. S. E. Embretson (Ed.). *Measuring Psychological Constructs: Advances in Model-Based Approaches*. American Psychological Association.

Wang, S., & Wang, T. (2001). Precision of Warm's weighted likelihood estimates for a polytomous model in computerized adaptive testing. *Applied Psychological Measurement, 25*, 317-331.

Warm, T. A. (1989). Weighted Likelihood Estimation of Ability in Item Response Theory. *Psychometrika*, *54*, 427–450.

Widaman, K. F., Ferrer, E., & Conger, R. D. (2010). Factorial invariance within longitudinal structural equation models: measuring the same construct across time. *Child Development Perspectives*, *4*(1), 10-18.

Wise, E. A. (2004). Methods for analyzing psychotherapy outcomes: A review of clinical significance, reliable change, and recommendations for future directions. *Journal of Personality Assessment, 82*, 50-59.

Woods, C. M. (2011). Ramsay-curve item response theory (RC-IRT) to detect and correct for nonnormal latent variables. *Psychological Measurement, 11*, 253-270.

Wright, A., Hannon, J., Hegedus, E. J., & Kavchak, A. E. (2012). Clinimetrics corner: a closer look at the minimal clinically important difference (MCID). *Journal of Manual and Manipulative Therapy, 20*(3), 160-166.

Zigmond, A. S., & Snaith, R. P. (1983). The hospital anxiety and depression scale. *Acta Psychiatrica Scandinavica, 67*(6), 361–370.

Acknowledgements

Firstly, I would like to thank my supervisors, Klaas and Wilco, for all their help during my PhD. I thank Klaas especially for his regular advice on writing. His timely and meticulous feedback has helped me a great deal in finishing the dissertation on time. And Wilco, without your input, completing it would have been impossible. I thank you for always being there when I needed help. I also want to thank the rest of the MTO crew for all the good times together. Eva, my dear office mate, you have been great during the four years we have worked together.

Moving to Tilburg after my Master's in Groningen was not easy as I had to leave all my friends behind. However, thanks to Olga and Sugi, I felt at home here from early on. Thank you for helping me settle down and find a place to stay in Tilburg. During later years, I also met other great people in Tilburg who made my days here fun and memorable. Filiz and Bilge, I thank you for being there in good and bad times. Your presence has definitely improved the quality of my life here in Tilburg. And Dino, my friend and my sports advisor, I want to thank you for teaching me all those fancy exercises and for simply being a nice person. Without you and Barend, aka my wingman and partner in crime, my Tilburg experience wouldn't have been the same.

Other people I want to thank are Gaby, Lissette, Diogo, Marijke, Ronald, Ufuk, Christina, Francesca, Byron et al., for the fun times together. And of course Goda, Roxi, Zina and the rest of the Groningen crew for being good friends. Even though it was not possible to meet you in person often, I still felt your positive presence in my life. The same holds for my friends from Azerbaijan. These are, in alphabetical order, Anar(s), Aqshin, Bahadir, Elkhan, Rashad, Rauf, Rovshan, Rufat, Ruslan H. and Tural.

Last but not least, I want to thank my dad, my mom and my sisters for all their support during my PhD and beyond. I have come this far thanks to their blessings and belief in me. I owe each one of you.